Contribution a l'analyse du langage oral spontané

Mohamed-Zakaria Kurdi

UNIVERSITE JOSEPH FOURIER-GRENOBLE1
INFORMATIQUE ET MATHEMATIQUE APPLIQUEE

THESE
pour obtenir le grade de

DOCTEUR DE L'UNIVERSITE JOSEPH FOURIER
(arrêtés ministériels du 5 juillet 1984 et du 30 mars 1992)

Discipline informatique

par

Mohamed-Zakaria KURDI

Le 18 avril 2003

Contribution à l'analyse du langage oral spontané

Jury :

Rapporteurs: Jean-Marie Pierrel
Gérard Sabah

Examinateurs: Jean-Yves Antoine
Christian Boitet (Président)
Alain Lecomte

Directeur de thèse: **Jean CAELEN**

Thèse préparée au sein du laboratoire de Communication Langagière
et Interaction Personne-Système – Fédération IMAG

Remerciements

Mes travaux de thèse présentés dans ce manuscrit n'auraient pu aboutir sans l'aide et la présence de nombreuses personnes que je tiens à remercier ici :

Tout d'abord, Jean CAELEN mon directeur de thèse pour m'avoir accueilli à GEOD, pour sa confiance, sa présence et son aide durant les années de préparation de cette thèse.

Gérard SABAH et Jean-Marie PIERREL qui m'ont fait le plaisir d'accepter la charge de rapporteur ainsi que pour leurs remarques constructives sur mon travail.

Christian BOITET pour les conseils chaleureux et les remarques critiques qu'il a su me prodiguer ainsi que pour l'honneur qu'il m'a fait en présidant le jury de cette thèse.

Jean-Yves ANTOINE pour s'être intéressé à mon travail dès le début, pour avoir animé le groupe de travail sur la compréhension robuste (qui était une excellente occasion pour rencontrer des thésards d'autres universités qui travaillent sur la même thématique) ainsi que pour sa participation à mon jury de thèse.

Alain Lecomte pour les différentes discussions enrichissantes qu'on a eu ainsi que pour avoir accepté de participer à mon jury de thèse.

Je tiens aussi à remercier tous les membres de l'équipe GEOD ainsi que ceux du NISLab à Odense. En particulier, j'aimerais remercier mes voisins de bureau Luis VILLASEÑOR-PINEDA et Mohamed AHAFHAF à GEOD ainsi que Mykola KOLODNYTSKY et Michel GENEREUX au NISlab.

Je remercie finalement Niels-Ole BERNSEN, directeur du NISLab, pour m'avoir accordé sa confiance et pour ses encouragements au cours de mon travail dans son laboratoire.

Table des matières

Introduction générale

L'interprétation de la parole est un processus qui met en œuvre des mécanismes très complexes et très divers afin d'analyser un énoncé. Une classification extrêmement simplificatrice du processus d'interprétation consiste à séparer le traitement de la parole en deux étapes distinctes : la reconnaissance et la compréhension. Selon cette distinction dichotomique, la reconnaissance consiste à identifier les phonèmes et à les assembler en mots. La compréhension est considérée comme étant le mécanisme selon lequel on associe une interprétation à l'énoncé reconnu, en prenant en considération le contexte dans lequel cet énoncé est émis. Malgré l'existence de différents travaux en psycholinguistique expérimentale (voir par exemple (Schwartz, 1996), (Kurdi, 1996) pour une revue générale de ces travaux) qui montrent que les relations entre la reconnaissance et la compréhension sont trop complexes pour être séparées de cette manière, cette distinction a été adoptée dans la majorité des travaux récents dans le domaine du traitement automatique du langage oral visant à simuler ce processus de compréhension chez les humains. Ainsi, on distingue entre deux champs de recherche au sein du domaine du traitement automatique du langage oral : la reconnaissance de la parole et l'analyse linguistique du langage oral qui correspondent approximativement à la perception et la compréhension chez l'humain. Dans cette thèse, notre travail s'inscrit dans le contexte des recherches sur l'analyse linguistique du langage oral. Ce domaine a connu récemment des avancées significatives grâce aux développements technologiques dans le domaine de l'Intelligence Artificielle (IA) en général, à l'amélioration de la qualité des systèmes de reconnaissance automatique de la parole, et à la proposition de modèles linguistiques plus fins qui sont aptes à décrire les différentes propriétés du langage oral (Cole, 1996).

Afin de construire une représentation sémantique correspondant à un énoncé quelconque, un système d'analyse linguistique du langage oral doit surmonter des obstacles dont les principaux sont les suivants :

- **Problème de la qualité de la reconnaissance de la parole :** les systèmes actuels de reconnaissance de la parole spontanée sont loin de donner des performances satisfaisantes. En effet, le taux de reconnaissance varie considérablement selon plusieurs facteurs, comme le débit de la parole, la quantité du bruit (la qualité de la reconnaissance baisse avec la diminution de rapport signal/bruit), etc. Ces erreurs consistent généralement en insertion, suppression ou substitution de certains mots de l'énoncé. Cela nécessite le recours à une approche très flexible afin de corriger le maximum de ces erreurs d'une part et de réduire l'effet des erreurs non corrigées sur l'interprétation de l'énoncé d'autre part.

- **Problème des spécificités grammaticales de la parole spontanée :** comme nous allons le voir en détail dans la première partie de cette thèse, la syntaxe de l'oral présente certaines spécificités qui nécessitent d'être prises en considération afin d'effectuer une analyse correcte des énoncés oraux.
- **Problèmes des extragrammaticalités de l'oral :** selon différentes études menées sur plusieurs corpus (Nakatani et Hirschberg, 1994), (Heeman, 1997), des phénomènes comme les répétitions, les autocorrections ou les faux-départs apparaissent dans environ 10% des énoncés d'un dialogue. Ces phénomènes nécessitent un traitement particulier afin d'éviter les erreurs d'analyse syntaxique et sémantique qu'ils peuvent causer.

Avec le développement des nouvelles technologies de la communication ainsi que des techniques de reconnaissance de la parole, on assiste à l'extension du cahier des charges des systèmes d'analyse linguistique du langage oral. Les exigences principales sont :

- **Augmentation de la finesse d'analyse :** cela nécessite l'utilisation de modèles linguistiques précis du langage oral .
- **Elargissement des domaines de dialogue :** cela contribue à l'augmentation du nombre des concepts et des mots dans le dialogue et par conséquent à l'augmentation de l'ambiguïté sémantique et lexicale.
- **Conditions réelles d'utilisation :** cela implique une couverture syntaxique de l'oral plus large ainsi que la prise en compte des extragrammaticalités.

Ces nouvelles exigences mettent les concepteurs de systèmes d'analyse du langage oral devant le dilemme suivant[1] :

- Pour répondre à la condition de finesse de l'analyse, les chercheurs ont souvent recours aux formalismes syntaxiques classiques couplés à une approche d'analyse complète. Malgré ses avantages en terme de finesse, ce choix conduit directement à une baisse importante de la robustesse étant donné que les formalismes syntaxiques classiques ainsi que l'approche d'analyse complète ont été conçus initialement dans le contexte de l'analyse écrite et ne sont donc pas adaptés aux particularités grammaticales et extragrammaticales de l'oral ni au traitement d'énoncés ayant des erreurs de reconnaissance.
- Pour répondre à la condition de la robustesse, les chercheurs utilisent des approches d'analyse superficielles et descendantes basées principalement sur la sémantique et combinées à des

[1] Ce dilemme se voit clairement dans le projet Verbmobil (Wahlster, 2000). En effet, comme nous allons le voir en détail plus loin, les systèmes d'analyse superficielle ont été plus robustes que ceux qui donnent une analyse profonde.

approches d'analyse partielles ou à base de mots clés. Ces approches, malgré leur robustesse, ne permettent souvent pas de traiter correctement les énoncés linguistiquement complexes.

1 Objectif et contribution de la thèse

Dans cette thèse, nous proposons une approche qui optimise le rapport finesse-robustesse dans un système d'analyse du langage oral. Avant de montrer la contribution de notre travail par rapport à la problématique de notre thèse présentée ci-dessus, nous définissons à notre propre compte les notions de base de cette problématique :

1. **L'analyse linguistique :** par analyse linguistique nous entendons l'association d'une représentation formelle (syntaxique et/ou sémantique) à un énoncé isolé de son contexte dialogique. Nous avons préféré l'utilisation de cette expression (analyse linguistique) plutôt que le mot compréhension étant donné que la compréhension couvre des domaines qui relèvent du dialogue comme la résolution de l'anaphore ou de l'ambiguïté contextuelle qui sortent du cadre de notre étude[2].
2. **La robustesse :** nous définissons la robustesse comme la capacité du système à donner une analyse correcte quelles que soient les conditions dans lesquelles l'analyse est faite. Dans le contexte des systèmes d'analyse linguistique de la parole, cela signifie que le système doit être capable de donner une **interprétation correcte** même dans les cas où l'énoncé contient des erreurs de reconnaissance, des extragrammaticalités, une construction syntaxique particulière, etc.
3. **La profondeur :** ce que nous entendons par profondeur est la capacité du système à construire une représentation syntaxique et sémantique d'un énoncé quelles que soient sa forme et sa complexité linguistique (constructions relatives, ellipses, incises, etc.). Les représentations fournies doivent refléter fidèlement toutes les variations linguistiques qui

[2] La compréhension peut être vue comme une contextualisation dialogique de l'analyse linguistique. Par exemple un énoncé elliptique comme : *deux* est interprété par un module d'analyse linguistique en l'associant à une représentation comme: nombre(deux). Le module de compréhension prend cette représentation et l'ancre dans le contexte de la conversation en cours. Ainsi, si l'énoncé *oui* est précédé par une question du système comme : *combien de chambres voulez vous*, le module de compréhension enrichit la représentation initiale obtenue avec le module d'analyse linguistique et sa sortie peut être une représentation comme : chambres_demandées(nombre (2)). Par extension, la compréhension peut être considérée comme la construction de l'analyse linguistique ainsi que sa contextualisation.

ont un effet sur le sens utile de l'énoncé. Le sens utile est le sens nécessaire pour le déroulement d'un échange dialogique pertinent entre deux agents.

Nous pouvons résumer les objectifs de notre thèse par les points suivants :

1. Etude des phénomènes extragrammaticaux en particulier en ce qui concerne leur régularité aussi bien que leur rapport avec la grammaire de la langue en général.
2. Proposition du formalisme Semantic Tree Association Grammar (Sm-TAG) qui est destiné au traitement des phénomènes grammaticaux du langage oral.
3. Implantation de quatre systèmes basés sur nos études des extragrammaticalités ainsi que sur la Sm-TAG et évaluation de l'adaptation de ces systèmes au traitement des phénomènes grammaticaux et extragrammaticaux du langage oral.

2 Le plan de la thèse

1.1 La première partie

Cette partie s'articule autour de trois chapitres :

Le premier chapitre présente les différentes spécificités communicationnelles et linguistiques du langage oral. Nous allons en particulier, nous concentrer sur les aspects grammaticaux et extragrammaticaux du langage oral.

Le deuxième chapitre est consacré à la présentation de deux formalismes qui sont utilisés pour la représentation des phénomènes grammaticaux du langage oral.

Dans le troisième chapitre, nous allons présenter les principales approches pour l'analyse syntaxique robuste ainsi que pour le traitement des extragrammaticalités du langage oral.

1.2 La deuxième partie

Consacrée aux études théoriques que nous avons effectuées, cette partie, est composée de deux chapitres :

Dans le premier chapitre, nous allons décrire notre méthode d'analyse du *Trains Corpus* ainsi que les résultats de cette analyse.

Dans le deuxième chapitre, nous allons présenter la formalisation de la grammaire sémantique de substitution d'arbres (S-TSG) ainsi que les différentes propriétés formelles et linguistiques de la grammaire sémantique d'Association d'Arbres (Sm-TAG) que nous avons proposé spécifiquement pour prendre en considération les phénomènes de l'oral.

1.3 La troisième partie

Consacrée aux applications des modèles théoriques, cette partie contient deux chapitres :

Le premier chapitre sera consacré au système Corrector qui est une mise en œuvre applicative de notre modèle sur les extragrammaticalités. Le système Corrector est basé sur une approche intégrée qui combine des techniques diverses (notamment la reconnaissance de patron et l'analyse superficielle) pour traiter les différentes formes d'extragrammaticalités.

Consacré aux applications des formalismes S-TSG et Sm-TAG, le deuxième chapitre porte sur les systèmes SAFIR, OASIS et NAVIGATOR. Ces trois systèmes sont basés sur une approche d'analyse partielle et sélective qui leur permet d'être robustes par rapport aux différentes sources de problèmes d'analyse comme les erreurs de reconnaissance et les extragrammaticalités. Les méthodes utilisées pour l'évaluation de ces systèmes ainsi que les résultats obtenus seront aussi présentés dans ce chapitre.

Partie I : Le langage oral spontané, sa représentation grammaticale et son analyse automatique

0 Introduction de la première partie

Cette partie détaille les différentes propriétés linguistiques du langage oral ainsi que les principaux formalismes syntaxiques et sémantiques qui peuvent être utilisés pour représenter ces différentes propriétés. Ainsi, cette partie s'articule autour de deux chapitres :

- Les différentes spécificités linguistiques du langage oral. Dans cette présentation, une attention particulière sera accordée sur les phénomènes syntaxiques et discursifs observés à l'oral ainsi que sur les différents phénomènes d'extragrammaticalités.
- Les deux formalismes grammaticaux qui ont inspiré notre travail : LTAG et la grammaire sémantique.
- Les approches principales pour l'analyse robuste du langage oral spontané.

1 Chapitre I.1 : Spécificités linguistiques du langage oral

1.1 Introduction

Ce chapitre a pour objectif de montrer les différents aspects linguistiques du langage oral spontané oraux avec une mise en évidence des phénomènes grammaticaux, extragrammaticaux et discursifs.

1.2 La syntaxe du langage oral

Dans le domaine de la parole, un nombre assez considérable d'études a porté sur les aspects phonétique et phonologique, mais la syntaxe, qui est pourtant une discipline centrale dans la linguistique, est la seule à rester soumise au règne du scripturocentrisme comme le souligne (Kerbat-Orecchioni, 2001). En effet, les études syntaxiques ont essentiellement porté sur l'écrit en négligeant l'oral considéré comme une forme appauvrie et parfois déviante de l'écrit. Le manque de ressources linguistiques à cause des difficultés de collecte et de transcription de dialogues oraux (voir (Blanche-Benveniste, 1987) pour une revue générale de ces problèmes) ainsi que l'importance assez limitée du traitement syntaxique de l'oral avant les années quatre-vingt-dix constituent d'autres raisons à ce retard.

1.2.1 Les aspects syntaxiques de base

Nous allons présenter dans les paragraphes suivants les principaux aspects syntaxiques en français parlé classés par grandes classes de phénomènes.

2.1.1.1 La topologie en français parlé

Il s'agit de l'ordre selon lequel les mots sont agencés au sein de la phrase. En général, la topologie permet de savoir la fonction d'un argument selon sa position par rapport au verbe (Lazard, 1994). Par exemple, le français est une langue à ordre SVO (Sujet Verbe Objet). Selon les langues, cet ordre peut varier de fixe à totalement variable. A l'écrit, le français respecte parfaitement l'ordonnancement standard. Cependant, l'oral ne semble pas obéir à la même règle. Par exemple, les énoncés suivants sont parfaitement possibles dans une conversation parlée :

Mon cahier je l'ai oublié à la maison	(antéposition d'un SN : OSV)	(1)
A 200 mètres vous trouvez une pharmacie	(antéposition d'un SP : OSVO)	(2)
Moi mon père je l'aime beaucoup	(double marquage : SOSOV)	(3)

La question qui se pose est de savoir quelle est l'importance de ces cas en terme de fréquence dans les conversations parlées puis de savoir si cette fréquence dépend du contexte syntaxique (c'est-à-dire,

est-elle plus importante dans un contexte syntaxique C1 que dans un autre contexte syntaxique C2) (Antoine et Goulian, 2001) ont essayé de répondre à ces questions dans une étude récente basée sur trois corpus de français parlé[3]. Ainsi, ces chercheurs ont montré que dans des situations ordinaires le langage finalisé respecte l'ordonnancement privilégié.

2.1.1.2 L'accord en genre et en nombre

En français, il s'agit d'un mécanisme selon lequel un nom ou un pronom donné exerce une contrainte formelle sur les pronoms qui le représentent, sur les verbes dont il est sujet, sur les adjectifs ou participes passés qui se rapportent à lui (Dubois, 1994). Selon les constructions, l'accord est plus ou moins respecté à l'oral. Par exemple, le non-respect de l'accord entre le substantif et/ou ses adjectifs sont très rares (exemples 4 et 5), alors que l'accord en genre entre l'attribut et le mot auquel il se rapporte est très fréquent (Sauvageot, 1972). Voici une série d'exemple de non-respect de l'accord (les trois premiers sont tirés de (Sauvageot, 1972)).

Une **voiture émetteur** (4)

Les **revenus salariaux** (5)

Les **dispositions** que nous avons **pris** (6)

C'**est** mes **amis** (7)

Notons que même en cas de respect de l'accord, ce respect n'est souvent pas marqué par des réalisations phonétiques perceptibles par l'auditeur de l'énoncé. Par exemple, le *e* utilisé pour marquer le genre féminin n'est associé à un phonème que dans des contextes exceptionnels comme lorsqu'il est précédé d'un *s* : émise, admise.

1.2.2 Exemples de constructions syntaxiques complexes et leurs spécificités à l'oral

2.1.1.3 L'interrogation

L'interrogation sous toutes ses formes est un moyen linguistique particulièrement important dans le dialogue. Trois dispositifs sont utilisés en français parlé pour marquer l'interrogation (Gadet, 1989), (Capelle et Frérot, 1979) :

1. **L'inversion :** il s'agit de placer le verbe avant le sujet (qui peut être un nom ou un pronom) comme dans les exemples suivants :

   La chambre est-elle libre ? (8)

   Arrive-il ce soir ? (9)

   Ce dispositif est utilisé à la fois à l'oral et à l'écrit.

[3] Il s'agit des corpus : Air France (Morel, et al., 1989), Murol (Bessac et Caelen, 1995), Levelt (Ozkan, 1994).

2. **Les interrogatifs :** sont des pronoms, des adjectifs ou des adverbes qui indiquent l'interrogation sans changer l'ordre des éléments de l'énoncé. Par contre, les interrogatifs eux-mêmes peuvent venir au début (les exemples 10, 13, 15, 16 et 17) ou à la fin de l'énoncé (les exemples 11, 12 et 14).

Est-ce qu'elle est chère ? (adverbe d'interrogation) (10)

Elle est très chère **n'est-ce pas** ? (11)

C'est très loin d'ici **non** ? (12)

Laquelle des deux est moins chère ? (pronom interrogatif objet) (13)

Son prix c'est **combien**? (14)

Qu'est-ce que vous avez comme services ? (15)

C'est quand que commence le spectacle ? (16)

Quels services proposez-vous ? (adjectif interrogatif) (17)

Les exemples 11, 12 et 14 sont propres à l'oral. Pour le reste, il s'agit d'exemples partagés à l'oral et à l'écrit.

3. **L'intonation :** L'intonation montante marque toutes les formes d'interrogation celles qui impliquent un élément syntaxique ou pas. Dans ce deuxième cas, elle permet toute seule d'exprimer l'interrogation en maintenant généralement l'ordre des mots l'énoncé assertif. L'intonation interrogative est généralement montante (exemple 19) contrairement à l'intonation des énoncés déclaratifs qui est descendante (exemple 18).

Elle est confortable (énoncé déclaratif) (18)

Elle est confortable ? (énoncé interrogatif) (19)

L'interrogation peut porter sur la totalité de l'énoncé ou sur une partie seulement.

Les énoncés interrogatifs qui n'impliquent que la prosodie pour marquer l'interrogation constituent le cas le plus fréquent à l'oral (Gadet, 1989). Cette fréquence est due essentiellement à des raisons d'économie et d'efficacité. Les autres dispositifs de marquage de l'interrogation étant considérés comme redondants d'une part et d'autre part ces dispositifs ne sont pas nécessaires pour faciliter l'accès perceptif de l'auditeur à l'aspect interrogatif de l'énoncé : l'intonation étant très facilement perceptible par les auditeurs à cause notamment de sa durée importante en général.

2.1.1.4 Les relatives

La relative est considérée comme l'un des principaux exemples de divergence entre l'oral et l'écrit (Gadet, 1989). Une proposition relative est une proposition qui contient un pronom relatif enchâssé dans le syntagme nominal constituant d'une phrase dite principale. Le syntagme nominal qui sert de base à l'enchâssement est appelé antécédent. Soit les énoncés :

Le professeur dont je parle (20)

Le professeur de qui je parle (21)

Outre les formes dites *standards* utilisées à la fois à l'oral et à l'écrit (comme dans les exemples (20) et (21)), trois types de relatives peuvent être observés uniquement à l'oral (Gadet, 1989) :

- **Les relatives dites de *français populaire :*** ces relatives peuvent être réalisées avec un clitique (22), avec un groupe prépositionnel (23) ou avec un possessif (24) :

  Le prof que j'en parle (22)

  Le prof que je parle de lui (23)

  Le prof que je parle de sa matière (24)

- **Relative défective :** cette forme est dite défective parce qu'elle crée une ambiguïté entre l'objet direct et l'objet indirect c'est pourquoi elle est la moins fréquente comparée aux autres.

  Le mot (du directeur) **que** je parle (25)

- **Relative pléonastique :** ces relatives ont une structure similaire aux relatives dites de *français populaire* et elles s'en distinguent par le pronom d'objet.

  Le prof dont j'en parle (26)

  Le prof dont je parle de lui (27)

  Le prof dont je parle de sa matière (28)

1.3 Les extragrammaticalités du langage oral

1.3.1 Terminologie

Extragrammaticalités, inattendus structurels, spontanéités, non-continuités (disfluencies), autant de mots ont été proposés dans la littérature pour désigner les phénomènes spontanés de l'oral comme l'hésitation, la répétition, l'autocorrection, etc. Chacun de ces termes a sa motivation. Dans *Inattendus structurels,* d'une part, le mot *structurel* est trop général et peut désigner toute sorte de phénomènes linguistiques et d'autre part, le mot inattendu porte un jugement *a priori* sur la prédictibilité d'un de ces phénomènes. En effet, plusieurs études ont montré que, sachant le contexte, ces phénomènes sont parfaitement prévisibles (Lickley, 1994), (Shriberg, 1994). Quant au mot *spontanéité*, il est trop vague

et général et ne donne pas d'indication sur la nature des phénomènes désignés. En particulier ce terme ne permet pas de distinguer les spontanéités grammaticales des spontanéités extragrammaticales. Le terme *non-continuités* (disfluency) (Shriberg, 1994), (Lickley, 1994), (Heeman, 1997), (Core, 1999), porte essentiellement sur l'aspect phonétique des phénomènes. Or, les phénomènes de l'oral ne sont pas toujours accompagnés de variations phonétiques particulières ou d'interruptions comme le laisse entendre ce terme. Le terme extragrammaticalité (Carbonell, 1984) nous semble le plus approprié. En effet, il est suffisamment général et précis pour couvrir les différents phénomènes spontanés de l'oral qui ne dépendent pas directement de la syntaxe de la langue.

1.3.2 Le paradoxe des extragrammaticalités

La différence principale entre les phénomènes grammaticaux de l'oral et les extragrammaticalités est que les premiers dépendent entièrement de contraintes inhérentes à la grammaire (ils sont liés à la compétence linguistique) alors que les seconds sont liés à l'usage de la langue dans les conditions réelles (phénomènes de performance). Autrement dit, l'occurrence d'une extragrammaticalité dépend principalement de raisons externes à la langue. Ainsi, nous nous trouvons devant le paradoxe suivant : les extragrammaticalités, tout en étant causées par des raisons complètement externes à la grammaire de la langue, se manifestent sous une forme grammaticale : des constructions syntaxiques considérées comme étant mal-formées par rapport à la grammaire de la langue mais dont la construction n'est pas **complètement** indépendante d'elle (l'extragrammaticalité se manifeste comme une série d'items lexicaux, de syntagmes dont dépendent directement de la grammaire de la langue. Pour mettre au clair ce paradoxe, nous dressons un schéma général de l'émission des extragrammaticalités dans la figure suivante :

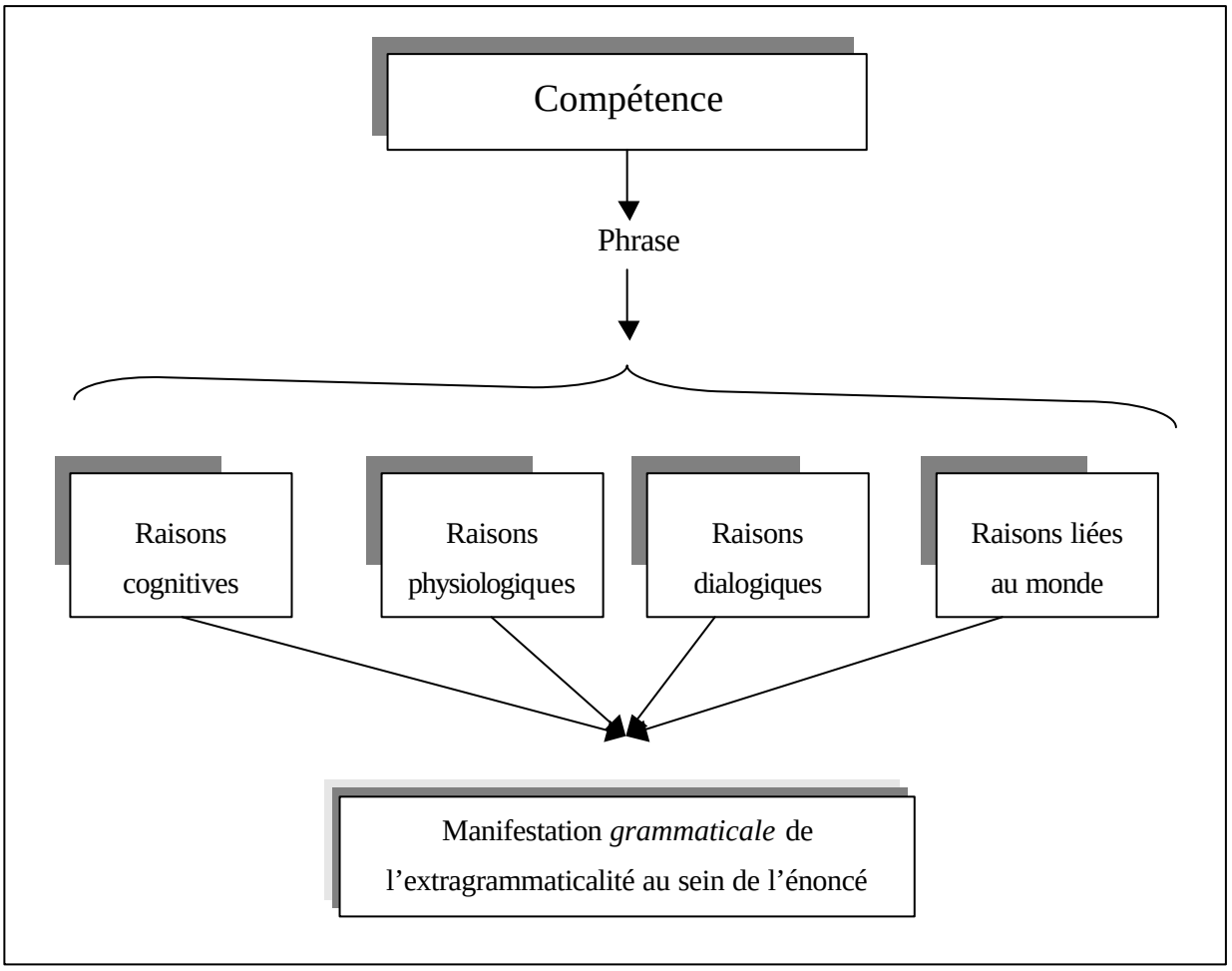


Figure 1. Schéma général des sources d'émission d'une extragrammaticalité

Comme nous pouvons le constater dans le schéma précédent, les raisons des extragrammaticalités sont multiples. En voici une description générale :

1. **Les raisons cognitives :** il s'agit d'un ensemble de raisons qui peuvent varier entre l'état émotionnel du sujet, son degré de concentration et la complexité de la tâche qu'il doit résoudre. Par exemple, une tâche difficile nécessite plus de calculs cognitifs qu'une tâche simple. Dans certains cas, cela peut déclencher des extragrammaticalités afin de remplir le silence nécessaire à la réflexion.
2. **Les raisons physiologiques :** il s'agit d'un ensemble de raisons qui sont liées principalement à la production sonore de la parole. Cela peut consister à adapter la segmentation de l'énoncé au rythme de la respiration dans les cas d'un effort physique important ou le besoin soudain de dégagement du conduit respiratoire qui se traduit par des toux volontaires ou involontaires et qui ont par effet l'interruption du flux de la parole.
3. **Raisons dialogiques :** certaines extragrammaticalités résultent de la négociation de la prise du tour de parole. Ainsi, dans le cas de réussite de la prise du tour de parole par un interlocuteur,

l'énoncé en cours d'émission est interrompu. Dans ce genre de situations, on assiste typiquement à des cas d'incomplétude. Pour ailleurs, nous avons observé informellement des tentatives non réussies de prise de la parole qui se sont traduites par une déconcentration du locuteur et son émission de certaines extragrammaticalités avant de pouvoir continuer son énoncé normalement.

4. **Raisons liées au monde :** il s'agit d'un nombre infini d'événements qui peuvent parfois capter de manière très forte l'attention du locuteur et qui ont pour résultat soit l'arrêt immédiat de la prononciation soit une déconcentration importante du locuteur. Par exemple, un conducteur en cours de conversation et dont la voiture est heurtée par un vélo ou une autre voiture peut arrêter son énoncé et en commencer un autre pour répondre à la situation urgente.

1.3.3 Le schéma général des extragrammaticalités

Généralement, les extragrammaticalités de l'oral peuvent être divisées en trois zones temporelles selon le schéma présenté dans la figure 2 (Shriberg, 1994).

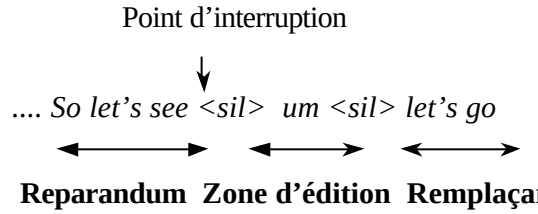


Figure 2. Schéma général des extragrammaticalités

1. **Le remplaçant :** Il s'agit de la zone que le locuteur introduit pour remplacer le reparandum.
2. **Le reparandum :** C'est la partie de l'énoncé que le locuteur juge incorrecte ou non nécessaire et qu'il remplace par le segment remplaçant.
3. **La zone d'édition :** C'est l'ensemble des mots qui séparent les deux zones précédentes et qui commence généralement par le point d'interruption, qui est caractérisé par une augmentation considérable de la fréquence fondamentale sur une hésitation ou un mot incomplet. Parfois la zone d'édition peut être porteuse de sens, comme c'est le cas généralement des expressions phatiques utilisées pour maintenir le contact et remplir le silence (*wait a moment please*).

Malgré son intérêt, ce schéma ne permet pas de rendre compte de tous les phénomènes d'extragrammaticalités en particulier les extragrammaticalités lexicales et les faux-départs.

Dans les paragraphes suivants nous allons présenter les différentes formes d'extragrammaticalités que nous avons classé dans deux groupes : les extragrammaticalités lexicales et les extragrammaticalités supralexicales.

1.3.4 Les extragrammaticalités lexicales (ELs)

Il s'agit d'un ensemble de phénomènes lexicaux propres à la langue parlée. Ces phénomènes peuvent avoir plusieurs formes et peuvent être de différentes natures : morphologique, phonétique.

2.1.1.5 Les pauses

La pause est l'un des phénomènes les plus caractéristiques de la production de la parole spontanée. D'un point de vue communicatif, en général, les pauses sont produites lorsque le locuteur veut se donner du temps pour préparer le reste de son énoncé. Deux types de pauses sont couramment observés à l'oral :

i. **Les pauses simples :** Les pauses simples sont caractérisées par l'absence totale du signal de parole pendant un laps de temps dont la durée varie selon les locuteurs (Lickley, 1994). La différence principale entre la pause simple et les silences normaux est que la pause ne semble dépendre d'aucune règle linguistique (une pause peut se produire même au sein d'un mot) alors que le silence est un phénomène acoustique dont la production dépend d'un ensemble de règles prosodiques et phonosyntaxiques assez complexes (Rossi, *et al.*, 1981). Par ailleurs, d'un point de vue fonctionnel, la pause a principalement pour fonction de préparer le reste de l'énoncé alors que les pauses ont des fonctions phonosyntaxiques précises comme la segmentation de l'énoncé.

ii. **Les pauses remplies (ou les hésitations) :** les pauses remplies, sont caractérisées par la continuation de production de signal acoustique pendant la période de pause (non-production de segments sémantiquement interprétables). Ainsi, l'hésitation sert à renforcer l'homogénéité discursive et à continuer à capter l'attention de l'interlocuteur même pendant les périodes de non-production d'unités linguistiques interprétables. Tout comme la pause simple, la pause remplie peut intervenir à tout moment dans l'énoncé sans nuire à son intelligibilité. Cependant, l'hésitation est plus dépendante du contexte que la pause simple. En effet, les expériences psycholinguistiques ont montré que les hésitations sont plus fréquentes devant des mots *lexicaux* (qui sont moins prédictibles) que devant les mots grammaticaux (Maclay et Osgood, 1967).

2.1.1.6 Les mots incomplets

L'incomplétude de certains mots de l'énoncé est un phénomène assez fréquent à l'oral : (…) I still have plenty of time and then <sil> <laughter> <sil> **thre-** <sil> and then it's <sil> **s-** four hours + back +

Bien qu'ils ne soient pas une fin en soi dans le traitement, les mots incomplets constituent un indicateur assez important pour la détection des extragrammaticalités supralexicales. Malheureusement, cette information n'est pas encore utilisable dans des conditions réelles, puisque les systèmes actuels de reconnaissance de la parole ne reproduisent pas les mots incomplets.

2.1.1.7 Les mots oraux

Il s'agit d'un ensemble de réalisations lexicales propres à l'oral et qui sont souvent des réalisations simplifiées de mots *standards*. Par exemple, *ouais* pour *oui*, *yeah* pour *yes*, etc. Les mots oraux eux-même constituent un phénomène grammatical normal lié principalement au niveau social de

l'utilisation de la langue : contexte formel ou informel. Cependant, dans certains contextes, ce niveau de la langue lui-même peut faire l'objet d'une correction. Par exemple, un locuteur qui juge que le mot oral qu'il a utilisé n'est pas approprié par rapport à la situation peut procéder au remplacement de ce mot par son équivalent formel comme dans l'énoncé : *ouais euh oui tout à fait*.

2.1.1.8 Les amalgames

Par amalgame nous voulons dire l'assemblage de deux mots ou plus dans une seule entité lexicale[4]. Ce genre d'assemblage est assez courant en français et en anglais parlés où l'on utilise souvent des formes lexicales pour désigner le sujet et le verbe en même temps comme : *Ch'ui* pour *je suis, I'd be* (I would be), *I'll* (I will), etc. Les amalgames sont des phénomènes grammaticaux dont l'utilisation dépend de contraintes sociales. Tout comme les mots oraux, les amalgames peuvent dans certians contextes être impliquées dans des extragrammaticalités visant à corriger le niveau de la langue.

1.3.5 Les Extragrammaticalités Supralexicales (ESLs)

Nous distinguons entre quatre phénomènes d'extragrammaticalités supralexicales : les répétitions, les autocorrections, les faux-départs et les incomplétudes.

2.1.1.9 Les répétitions

Il s'agit de la répétition d'un mot ou d'une série de mots. La répétition est définie sur des critères purement morphologiques. Par conséquent, la formulation et la paraphrase d'un énoncé ou d'un segment (où l'on répète deux segments qui ont le même sens) ne sont pas considérées comme étant des répétitions : (..) ce serait un vol Paris Delhi plus un vol un vol intérieur.

La répétition n'est pas toujours une redondance. Elle peut aussi avoir une fonction communicative. Par exemple, lorsqu'un locuteur n'est pas sûr que son message (ou une partie de son message) sera clairement perçu par son auditeur à cause d'une mauvaise articulation, d'un bruit dans le canal, etc. il le répète. Par ailleurs, la répétition est un moyen pragmatique assez fréquent pour marquer une affirmation ou une insistance comme dans l'énoncé 29 :

oui oui je vous en prends une (29)

Dans cet énoncé, la répétition du mot *oui* a une fonction d'affirmation.

2.1.1.10 Les autocorrections

L'autocorrection consiste à remplacer un mot ou une série de mots par d'autres afin de modifier ou corriger le sens de l'énoncé. L'autocorrection n'est pas complètement aléatoire et porte souvent sur un segment qui peut compter un ou plusieurs syntagmes (Core, 1999), c'est pourquoi elle est fréquemment accompagnée par une répétition partielle du segment corrigé. Soit l'énoncé 30 :

[4] En fait notre définition de l'amalgame ne couvre pas les phénomènes d'assemblage de morphèmes (comme au : à + le) qui sont commun à l'oral et à l'écrit.

Oui : j'ai **la** j'ai **les** pages Web oui (30)

Dan cet énoncé, l'autocorrection se fait en répétant le segment *j'ai* et en remplaçant le mot *la par*le mot *les*. On note que les deux mots ont la même catégorie morphologique (article défini) et la même fonction syntaxique (déterminant).

2.1.1.11 Les faux-départs

Il s'agit de l'abandon de ce qui a été dit et du recommencement d'un autre énoncé. Syntaxiquement, cela se manifeste par la succession d'un segment incomplet (ou mal formé) et d'un segment complet. Prenons l'énoncé :

(…) oui **c'est à** e ça se prend au deuxième étage (31)

Contrairement à l'autocorrection, il n'existe aucune analogie entre le segment remplacé et le reste de l'énoncé. Ainsi, nous pouvons remarquer dans l'exemple (31) que le segment abandonné *c'est à* n'a pratiquement pas de relation avec *ça se prend*…cette forme d'extragrammaticalité est la plus difficile à traiter étant donné que les critères de détection (essentiellement l'incomplétude d'un segment) sont très vagues et peuvent mener à de nombreux problèmes à la fois de surgénération et de sous-génération.

2.1.1.12 Les incomplétudes

Sur le plan syntaxique, un énoncé incomplet est un énoncé qui nécessite un ou plusieurs éléments à sa fin afin qu'il soit grammaticalement bien formé (au sens de la grammaire classique du terme) ou complètement interprétable sémantiquement. Plus concrètement, nous pouvons distinguer deux types d'énoncés incomplets :

1. Un énoncé auquel il manque un ou plusieurs constituants. Par exemple, l'énoncé (32) est considéré comme incomplet puisqu'il se termine par une conjonction de coordination qui nécessite l'existence d'une construction syntaxique qui complète l'énoncé.

   à peu près trois heures si vous devez changer à Vérone **et** (32)

2. Un énoncé dont tous les constituants nécessaires sont présents mais dont le dernier est incomplet. L'énoncé 33, par exemple, est considéré comme incomplet puisqu'il se termine par un constituant incomplet (un syntagme nominal dans ce cas).

   (…) et ils offrent des forf (33)

L'incomplétude est le phénomène le moins étudié parmi les différentes formes d'extragrammaticalités que nous avons passé en revue. En effet, il n'a pas été considéré par les principales études menées sur la détection e la correction des extragrammaticalités (Heeman, 1997), (Shriberg, 1994), (Core, 1999).

1.4 Les phénomènes discursifs observés dans le dialogue oral

Afin de lier les énoncés d'un dialogue les uns aux autres d'une part et d'autre afin d'ancrer ces énoncés au contexte dialogique, nous observons dans les dialogues oraux le recours à une série de dispositifs dont les principaux seront présentés dans les paragraphes suivants.

1.4.1 L'anaphore

L'anaphore est un moyen très important pour assurer le lien entre les différentes unités discursives tant à l'oral qu'à l'écrit. Son rôle est cependant plus central à l'oral qu'à l'écrit vu la structure dialogique qui implique un échange entre deux interlocuteurs et nécessite ainsi la référence à des parties du discours citées précédemment.

La définition généralement donnée de l'anaphore est la suivante : l'anaphore est un dispositif qui met en relation deux unités linguistiques dont la première est généralement pronominale (pronom personnel ou démonstratif) appelée anaphorique et dont la deuxième est un segment antérieur (souvent un syntagme nominal) comme dans l'exemple suivant (Dubois, 1994) :

Pierre, je le vois souvent. (34)

Une définition plus fine pour l'anaphore a été proposée dans (Krahmer et Piwek, 2000). Cette définition est basée sur plusieurs critères comme la dépendance contextuelle pour interpréter l'anaphore, le type de l'antécédent, le type de la relation entre l'anaphorique et l'antécédent, et l'intervalle des interprétations autorisées par l'anaphore.

Ainsi, dans un dialogue, nous pouvons distinguer deux types d'antécédents :

1. **Antécédent immédiat :** il s'agit des cas où l'anaphorique et l'antécédent se trouvent dans le même tour de parole, comme dans l'énoncé 38.
2. **Antécédent lointain :** ce sont les cas où l'anaphorique et l'antécédent se trouvent dans deux tours de parole différents et qui peuvent appartenir à deux locuteurs différents. Prenons comme exemple le segment suivant extrait du dialogue (jfs5.l) du corpus de réservation hôtelière :

   H= Alors, j'aurais **une chambre** pour une personne avec douche et WC au quatrième étage donnant sur le jardin à 380 Francs petit déjeuner compris.

   C= C'est très bien, je **la** prends. (35)

   Dans cet exemple, nous pouvons voir que l'anaphorique *la* et son antécédent se trouvent dans deux tours de parole de deux locuteurs différents (respectivement celui du client *C* et celui de l'hôtelier *H*). Notons aussi l'ambiguïté formelle de rattachement de l'anaphorique puisque dans l'énoncé *H* plusieurs syntagmes nominaux de genre féminin sont candidats : une chambre, une personne, douche.

Des cas encore plus complexes peuvent être observés où l'anaphorique se propage à travers plusieurs tours de parole. Dans ces cas, l'anaphore est très difficile à détecter puisqu'elle nécessite la

considération d'une fenêtre contextuelle très importante et qui contient souvent beaucoup d'ambiguïtés.

1.4.2 Les ellipses

L'ellipse consiste à omettre un certain nombre d'éléments d'un énoncé sans affecter son intelligibilité. En effet, l'omission crée un effet de puzzle qui permet à l'auditeur de retrouver les éléments omis et de compléter l'information. Tout comme l'anaphore, l'ellipse est un phénomène linguistique commun entre l'oral et l'écrit même si elle joue un rôle plus important à l'oral notamment dans les réponses à certaines questions. En général, l'ellipse constitue un moyen important pour éviter les redondances et par conséquent rendre la conversation plus simple et spontanée. Deux types d'ellipses peuvent être distingués :

2.1.1.13 Les ellipses situationnelles

Il s'agit d'un ensemble d'ellipses dont l'interprétation dépend étroitement de la situation d'élocution. Comme nous avons vu dans les paragraphes précédents, cette situation peut-être l'historique du dialogue, le contexte physique dans lequel se déroule la conversation, les connaissances générales du monde, etc. Voici un exemple d'ellipse situationnelle :

A : vous voulez une chambre simple ou une chambre double.

B : une simple. (36)

Dans cet exemple, nous remarquons la double ellipse dans la réponse : suppression de la formule de demande *je voudrais* et du mot *chambre* utilisée pour éviter la répétition de l'information fournie dans question posée.

2.1.1.14 Les ellipses grammaticales

Ce sont des ellipses qui consistent à omettre des mots que la connaissance syntaxique de la langue permet d'inférer. La forme la plus étudiée des ellipses grammaticales est l'ellipse verbale (Hardt, 1997). Dans ce genre d'ellipse, le syntagme verbal est supprimé dans des contextes où il est considéré inférable comme dans l'exemple :

Pierre mange des cerises, Paul des fraises. (37)

Dans l'énoncé précédent, le verbe de la deuxième proposition est supprimé, ce qui laisse entendre qu'il s'agit du même verbe que celui de la première proposition *mange*.

Par ailleurs, des ellipses mixtes peuvent être observées dans certains contextes. Pour illustrer ces ellipses, prenons comme exemple l'échange suivant :

A : Qu'est ce que tu en penses ? (38)

B : complètement d'accord. (39)

Dans l'énoncé (43) on a supprimé le segment *je suis* dont l'inférence est facile à partir de la règle syntaxique : sujet + verbe être + qualificatif, '*d'accord*'. Notons que la syntaxe toute seule est suffisante pour inférer le verbe être. La syntaxe a aussi joué un rôle direct dans l'inférence du sujet, cependant la forme du sujet (nom, pronom) ainsi que la personne (1ère du singulier, 2ème du pluriel, etc.) nécessite le contexte discursif. Ainsi, l'analyse finale de cette ellipse mobilise à la fois des connaissances syntaxiques et contextuelles.

Certaines formes de l'ellipse peuvent être vues comme un cas particulier de l'anaphore (Krahmer et Piwek, 2000). En effet, l'ellipse est basée sur un lien fort à une partie précédente du discours tout comme l'anaphore. Cependant, contrairement à l'anaphore où l'on a besoin d'un dispositif linguistique pour renvoyer à la partie précédente du discours, l'ellipse est caractérisée par la suppression des éléments communs avec ce qui a été dit.

1.4.3 Les déictiques (embrayeurs)

Il s'agit d'une classe de mots qui n'ont pas de référence propre dans la langue mais qui ne reçoive un sens que lorsqu'ils sont inclus dans un message. Les déictiques regroupent un ensemble relativement considérable de catégories grammaticales comme les démonstratifs, les adverbes de lieu et de temps, les pronoms personnels et les articles (Dubois, 1994).

Les déictiques peuvent faire référence à plusieurs aspects du contexte d'élocution comme :

1. **L'espace** dans lequel cet énoncé est produit. Exemple : tu peux le poser **ici** (représentation de l'espace).
2. **Le temps** au moment de l'énoncé. Exemple : il fait beau **aujourd'hui.**
3. **Le sujet parlant** (modalisation). Exemple : **je** le **lui** ai dit.

2 Chapitre I.2 : Les formalismes pour la représentation grammaticale du langage oral

Nous allons consacrer ce chapitre à la présentation de deux formalismes grammaticaux que nous avons jugé particulièrement pertinents pour notre travail. Il s'agit de la Grammaire d'Arbres Adjoints Lexicalisés et de la grammaire sémantique. Les motivations de notre choix sont le fait que les deux formalismes sont considérés comme standards respectivement dans le traitement de l'écrit et de l'oral. De plus, ces deux formalismes constituent les deux principales sources d'inspiration de notre formalisme Sm-TAG.

2.1 La Grammaire d'Arbres Adjoints Lexicalisés (LTAG)[5]

Le formalisme des Grammaires d'Arbres Adjoints a été décrit tout d'abord dans (Joshi et al., 75), sous le nom initial de *Tree Adjunct Grammar*. Ce formalisme a été ensuite développé par d'autres chercheurs particulièrement aux universités de Pennsylvanie, USA et Paris 7 (Voir (Abeillé, 1993) pour les étapes de développement de ce formalisme).

2.1.1 Définition formelle

D'un point de vue formel, le formalisme LTAG peut être défini comme un quintuplet (Σ, NT, I, A, S), où (Joshi et Schabes, 1999) :

i- Σ est un ensemble fini de symboles terminaux.

ii- *NT* est un ensemble fini de symboles non-terminaux. $\Sigma \cap NT = \phi$.

iii- *S* est le symbole non-terminal distingué : $S \in NT$.

iv- *I* est un ensemble fini d'arbres appelés arbres initiaux qui sont caractérisés par les points suivants :

- Les nœuds internes sont étiquetés avec des symboles non-terminaux.
- Les nœuds frontières des arbres initiaux sont étiquetés par des terminaux et des non-terminaux.

v- *A* est un ensemble fini d'arbres appelés arbres auxiliaires qui sont caractérisés par les points suivants :

- Les nœuds internes sont étiquetés avec des symboles non-terminaux.

[5] Lexicalized Tree Adjoining Grammar.

- Les nœuds sur les frontières des arbres auxiliaires sont étiquetés avec des symboles non-terminaux.

D'un point de vue fonctionnel, LTAG peut être décrit selon trois points :

1. Les unités de traitement (les arbres élémentaires).
2. Les opérations de composition.
3. Les traits et l'unification.

2.1.2 Les arbres élémentaires

Contrairement aux formalismes syntaxiques classiques basés sur le mot, l'unité de traitement dans une grammaire LTAG est l'arbre élémentaire. Ainsi, une grammaire LTAG peut être considérée comme un ensemble fini d'arbres élémentaires. Tout arbre élémentaire a au moins un de ses nœuds feuilles occupé par un item lexical qui joue le rôle de tête et qu'on appelle généralement l'ancre de cet arbre. En LTAG, la profondeur des arbres élémentaires n'est pas limitée à une branche[6]. Par ailleurs, deux types d'arbres élémentaires se distinguent dans ce formalisme :

2.1.1.15 Les arbres initiaux

Il s'agit d'un ensemble d'arbres qui se combinent par substitution et qui correspondent aux structures syntaxiques de base. Ces arbres sont généralement notés par (α).

2.1.1.16 Les arbres auxiliaires

Les arbres auxiliaires se combinent par adjonction. Ces arbres ont un nœud feuille (appelé nœud pied) portant un non-terminal de même catégorie que le nœud racine. Les arbres auxiliaires sont utilisés pour la représentation des modifieurs (adjectifs, adverbes, relatives), des verbes à complétives, des verbes modaux et des verbes auxiliaires. Ces arbres sont notés généralement par (β).

Les nœuds feuilles des arbres élémentaires peuvent être annotés par des symboles terminaux et non-terminaux. Deux types de nœuds annotés par des non-terminaux peuvent être distingués : les nœuds à substitution marqués par ($\downarrow$) et les nœuds à adjonction marqués par (*).

2.1.1.17 Contraintes de bonne formation des arbres élémentaires

La construction des arbres élémentaires obéit à quatre de principes de bonne formation (Abeillé, 1993).

1. **Principe d'ancrage lexical :** chaque arbre élémentaire doit être associé à au moins une tête lexicale. A la différence de HPSG[7] et d'autres formalismes, la tête lexicale d'un arbre élémentaire dans LTAG ne peut pas être vide[8]. De plus, un arbre élémentaire peut être ancré par un ensemble d'items lexicaux, on parle alors de co-têtes. Les co-têtes sont généralement des

[6] La profondeur est le nombre de branches qui séparent le nœud racine de l'arbre de l'ancre de cet arbre.

[7] Head Driven Phrase Structure Grammar.

[8] C'est l'une des raisons principales de la difficulté de traitement des ellipses dans le cadre de ce formalisme.

complémenteurs fonctionnels tel que *de* et *que*. Ainsi, à chaque entrée lexicale sont associés l'ensemble des structures qui caractériseront ses emplois possibles. Lexique et grammaire se confondent alors en un lexique syntaxique. D'un point de vue informatique, la lexicalisation permet de mobiliser uniquement le sous-ensemble des arbres élémentaires de la grammaire effectivement ancrés par les mots de la phrase.

2. **Principe de cooccurrence prédicat-arguments :** tout prédicat doit contenir dans sa structure élémentaire au moins un nœud pour les arguments qu'il sous-catégorise.
3. **Principe de consistance sémantique :** tout arbre élémentaire correspond à une représentation sémantique non vide.
4. **Principe de non compositionnalité :** un arbre élémentaire correspond à une seule unité sémantique.

Les principes sémantiques (2 et 3) sont assez vagues (aucune définition claire n'est donnée de ce qu'on entend par unité sémantique) et sont utilisés dans la LTAG essentiellement pour empêcher la plupart des éléments fonctionnels (prépositions, complémenteurs, etc.) de constituer des arbres élémentaires autonomes (principes 2). Le principe (3) sert à limiter la taille des arbres élémentaires et à empêcher l'ancrage de certains arbres par des éléments non nécessaires.

Voici quelques exemples d'arbres élémentaires en LTAG :

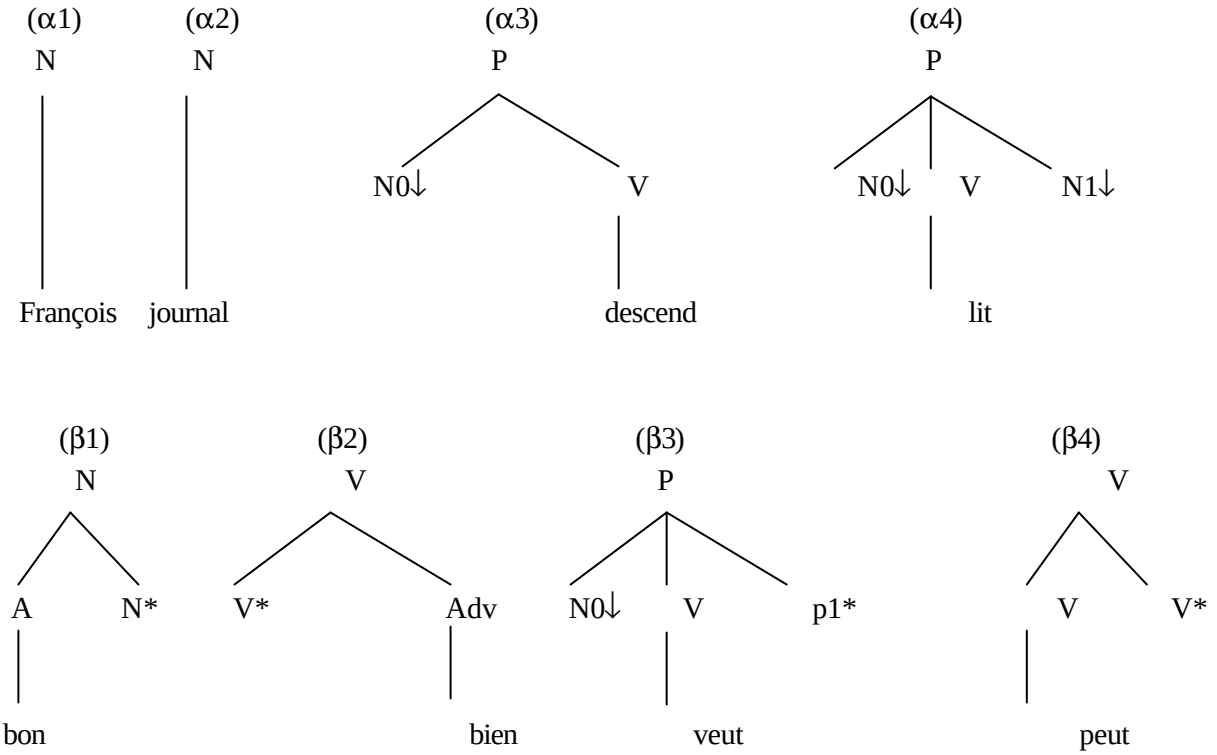


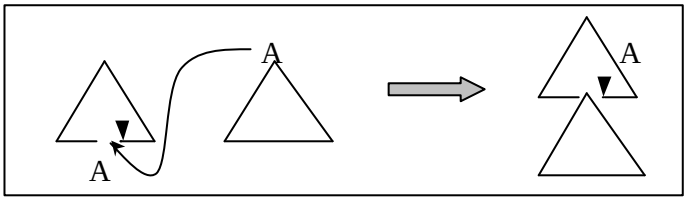
Figure 3. Exemples d'arbres élémentaires initiaux et auxiliaires

2.1.1.18 Les opérations de composition des arbres

Nous pouvons distinguer entre deux types de contraintes sur la composition des arbres élémentaires au sein du formalisme LTAG : les contraintes syntaxiques et les contraintes sémantiques. Ces différentes contraintes influencent la nature des opérations de composition utilisée. Dans la LTAG, deux opérations de composition syntaxique sont possibles : la substitution et l'adjonction.

2.1.2.1.1 La substitution

La substitution est similaire à l'opération de réécriture pour une CFG. Elle permet d'insérer un arbre, initial ou dérivé, à un nœud de substitution d'un arbre élémentaire ou dérivé qui est noté par le signe : ↓. La substitution est une opération obligatoire à un nœud terminal de substitution. Un exemple de substitution est l'insertion de l'arbre initial d'un déterminent dans l'arbre d'un groupe nominal.



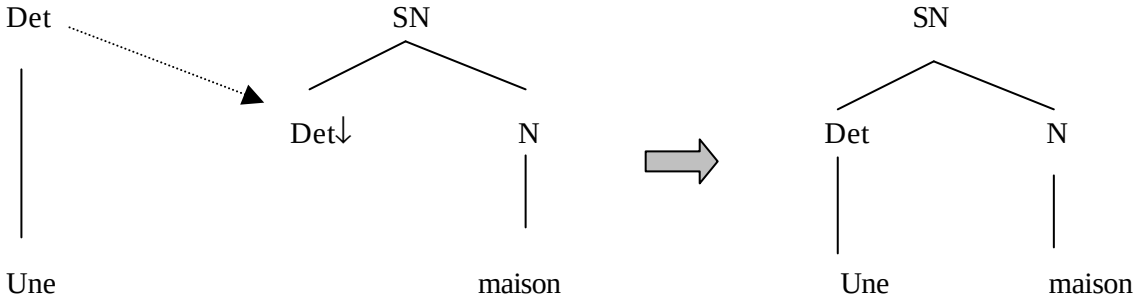


Figure 4. L'opération de substitution en LTAG

2.1.2.1.2 L'adjonction

L'adjonction est une opération spécifique au formalisme LTAG. Elle permet d'insérer un arbre auxiliaire (ou dérivé d'un auxiliaire) à un nœud interne ou racine d'un arbre élémentaire ou dérivé. Le nœud *X*, où a lieu l'adjonction, est remplacé par un arbre élémentaire dont la racine et le nœud pied doivent être étiquetés par la catégorie *X*. Le schéma général de l'opération d'adjonction est présenté dans la figure suivante.

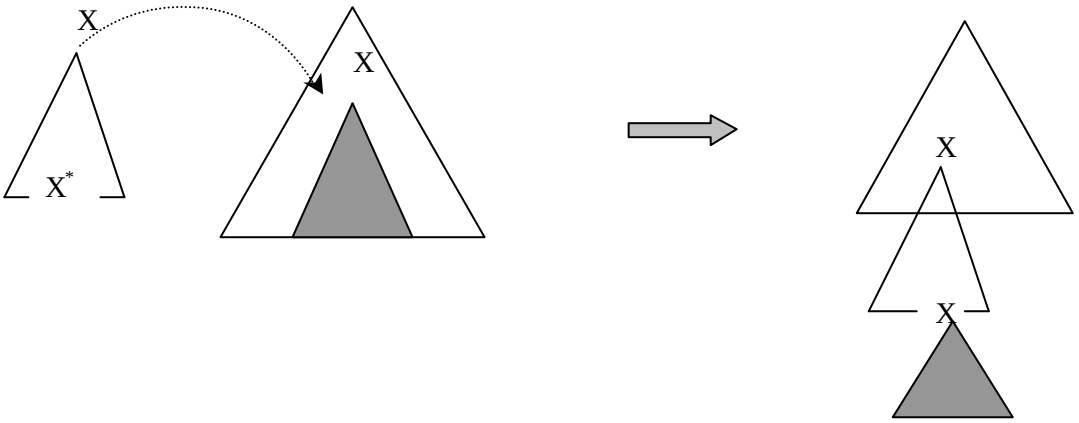


Figure 5. Schéma général de l'opération d'adjonction

Pour illustrer l'opération d'adjonction, prenons comme exemple l'insertion de l'arbre auxiliaire correspondant à l'adverbe au nœud intérieur V de l'arbre initial du verbe *marche*.

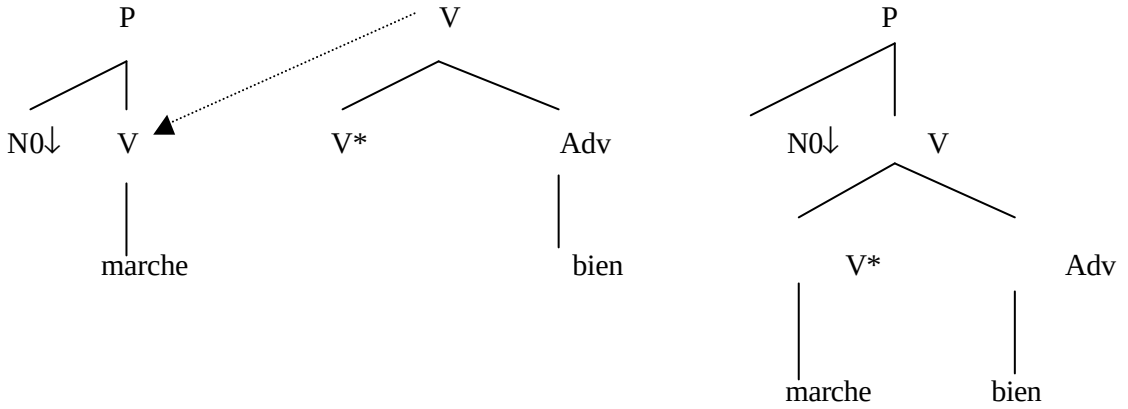


Figure 6. Exemple d'adjonction

Afin de contrôler l'adjonction dans une grammaire LTAG G = (∑, NT, I, A, S), trois types de contraintes sont définis sur l'adjonction à un nœud donné d'adjonction (Joshi et Schabes, 1999) :

- Adjonction sélective (SA (T))[9] : cette contrainte autorise l'adjonction aux seuls membres de l'ensemble T ⊆ A des arbres auxiliaires. Dans ce cas l'adjonction n'est pas obligatoire.
- Adjonction nulle (NA)[10] : cette contrainte interdit tout type d'adjonction au nœud donné.

[9] Simplification de *Selective adjunction of T*.

[10] Simplification de *Null adjunction*.

- Adjonction obligatoire (OA(T))[11] : cette contrainte oblige tout arbre auxiliaire membre de l'ensemble $T \subseteq A$ de s'adjoindre au nœud donné.

2.1.1.19 Spécificités de la composition syntaxique des arbres dans LTAG

Dans le formalisme LTAG, le processus de composition des unités élémentaires en unités plus larges ou la dérivation présente plusieurs spécificité comparé aux autres formalismes syntaxiques classiques. En effet, contrairement aux grammaires syntagmatiques de type CFG ou autre, la dérivation ne se caractérise pas comme une chaîne obtenue par d'autres chaînes mais comme un arbre obtenu d'autres arbres. Le résultat direct de cette différence est la distinction au sein du formalisme LTAG de deux modes de représentation du résultat de la dérivation qui sont l'arbre dérivé et l'arbre de dérivation.

1. **L'arbre dérivé :** est similaire à l'arbre d'analyse dans les formalismes syntagmatiques. Il s'agit d'un arbre à la racine duquel se trouve le symbole distingué du formalisme et aux feuilles duquel se trouvent les items lexicaux de l'énoncé analysé.
2. **L'arbre de dérivation :** ce genre d'arbres n'existe pas dans les formalismes syntagmatiques (nous pouvons dire que l'arbre de dérivation et l'arbre dérivé sont identiques dans ce genre de formalismes). Il s'agit d'un arbre dans lequel les nœuds portent des couples (arbre élémentaire, adresse du nœud de l'arbre supérieur où cet arbre a été inséré). La fonction principale des arbres de dérivation est de faire apparaître les dépendances entre les items lexicaux (tête des arbres élémentaires).

Voici un exemple d'un arbre de dérivation et d'un arbre de dérivation correspondant.

Arbre dérivé : Arbre de dérivation

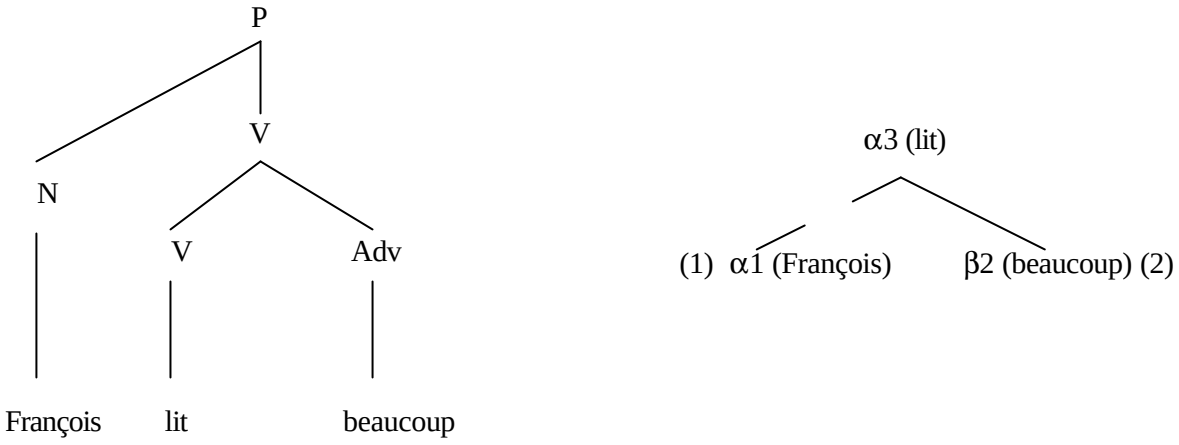


Figure 7. Un exemple d'un arbre dérivé et un arbre de dérivation correspondant

2.1.3 La composition sémantique et l'opération d'unification

Afin d'intégrer des contraintes sémantiques sur la composition des arbres des élémentaires dans le formalisme LTAG, les nœuds de ces arbres ont été décorés avec des structures de traits. Il s'agit de

[11] Simplification de *Obligatory adjunction of T.*

structures atomiques qui ont la forme (attribut, valeur). En TAG, les traits peuvent être morphologiques, syntaxiques et sémantiques. Les traits sont définis au niveau des arbres élémentaires et doivent être conservés dans les arbres dérivés. Deux types de traits sont associés à chaque nœud :

- Des traits amont (top) qui indiquent les relations du nœud avec les nœuds qui le dominent.
- Des traits aval (bottom) qui indiquent les relations du nœud avec les nœuds qu'il domine.

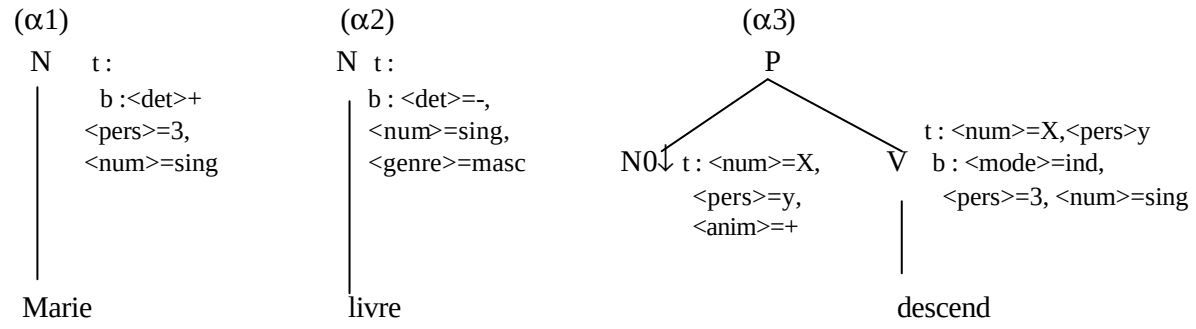


Figure 8. Exemples de structures de traits associés aux arbres élémentaires

Outre le regroupement des traits, l'opération d'unification permet d'exprimer les contraintes sur les rattachements possibles d'arbres. Ainsi, les deux opérations syntaxiques de formalisme TAG sont contraintes par l'unification de deux manières :

- En cas de substitution, les traits amont du nœud racine de l'arbre substitué doivent s'unifier avec les traits du nœud où il y a eu substitution.
- En cas d'adjonction, on doit avoir d'une part, unification des traits amont du nœud racine de l'arbre auxiliaire avec les traits amont du nœud recevant l'adjonction, et d'autre part, unification des traits du nœud pied de l'arbre auxiliaire avec les traits pied du nœud recevant l'adjonction.

A la fin d'une analyse, pour chaque dérivation complète obtenue, les parties amont et avale doivent s'unifier à chaque nœud de l'arbre dérivé correspondant. Voici un exemple d'unification :

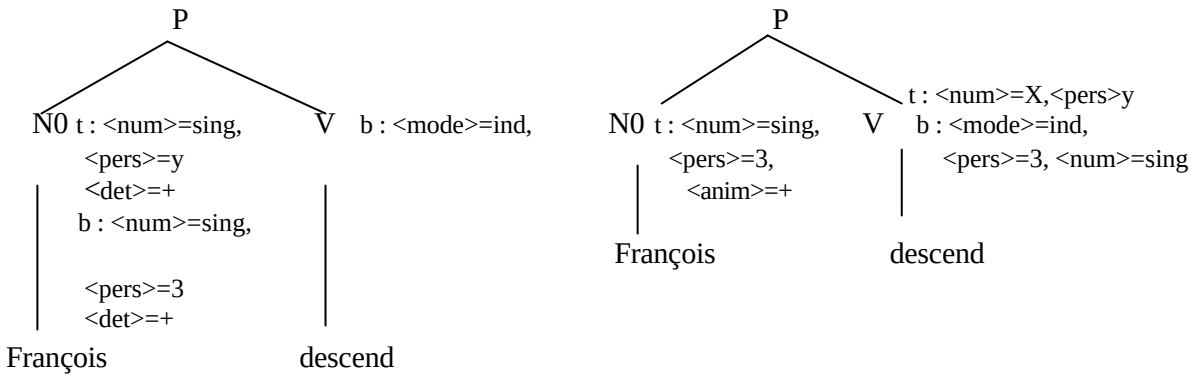


Figure 9. Un exemple d'unification

Le schéma de l'unification des traits en cas d'adjonction est présenté dans la figure suivante :

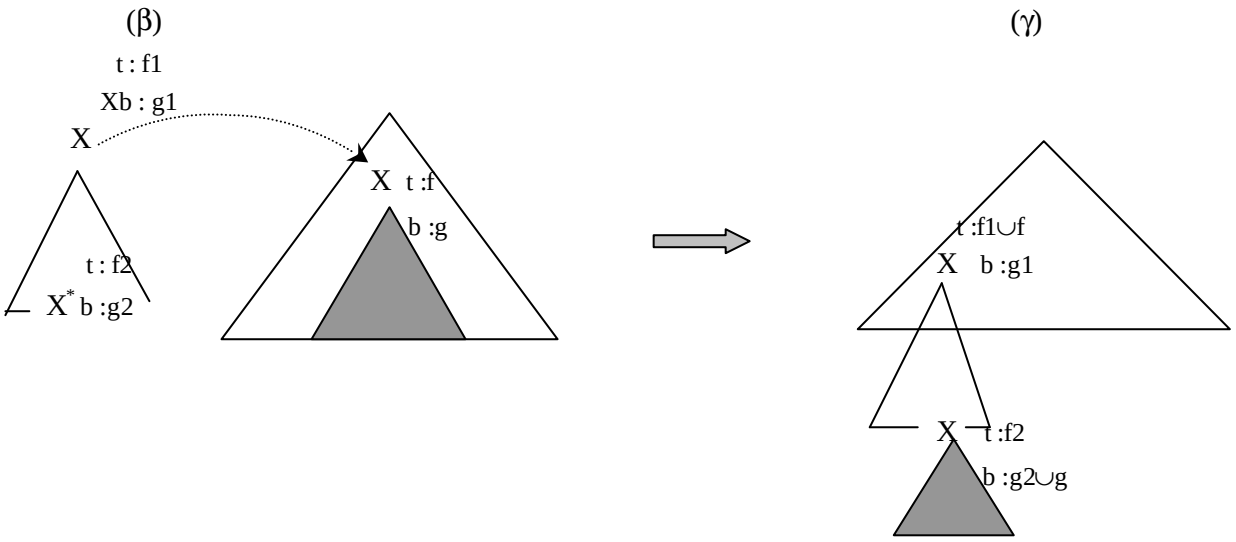


Figure 10. Schéma de l'adjonction avec unification

Malgré leur utilité dans le traitement, l'enrichissement du formalisme par des traits est une tâche assez difficile et nécessite beaucoup de travail. En ce qui l'adaptation au traitement des dialogues oraux, certains de ces traits semblent redondants et répétitifs. En effet, la connaissance du contexte, accessible via le modèle de la tâche, permet d'inférer une bonne partie de ces traits sans chercher à les vérifier de manière linguistique à travers les traits. Par exemple les traits relatifs au locuteur sont connus à priori à l'aide du modèle de la tâche : on sait qu'il s'agit d'un être humain singulier. On peut même savoir plus d'informations sur lui comme son rôle dans la conversation (cela dépend de la nature de l'application : client, expert cherchant à vérifier une information, chauffeur de voiture, etc.). De plus, l'inférence des informations à travers les traits n'est pas un processus fiable notamment à cause des erreurs de reconnaissance de la parole, des phénomènes linguistiques de l'oral, etc.

2.1.4 Les extensions du formalisme LTAG

Au cours de la dernière décennie, le formalisme LTAG a suscité un grand intérêt au sein de la communauté de linguistique computationnelle. Ainsi, différents sous formalismes inspirés des LTAGs ont vu le jour. Certains sont motivés par des raisons linguistiques comme la simplification de l'interaction syntaxe-sémantique (les TAGs Synchrones (Shieber et Schabes, 1990)), d'autres par des intérêts formels et computationnels comme la grammaire d'insertion d'arbres TIG[12] (Schabes, 1995). Dans les paragraphes suivants, nous nous contenterons de présenter les sous formalismes ayant un rapprochement direct avec notre travail.

[12] Tree Insertion Grammar.

2.1.1.20 Les TAGs Synchrones

Pour rendre l'interaction syntaxe-sémantique plus explicite au sein du formalisme LTAG, (Shieber et Schabes, 1990) ont proposé de paralléliser la structure syntaxique, représentée par les arbres élémentaires, et une structure de prédicat argument qui sert de squelette d'interprétation sémantique à l'arbre élémentaire auquel elle est associée. La représentation sémantique, elle aussi, a la forme d'une structure arborescente. Ainsi, à chaque arbre élémentaire est associé au moins un arbre sémantique et on définit des les liens entre les nœuds des deux arbres qui exercent des contraintes sur les dérivations possibles.

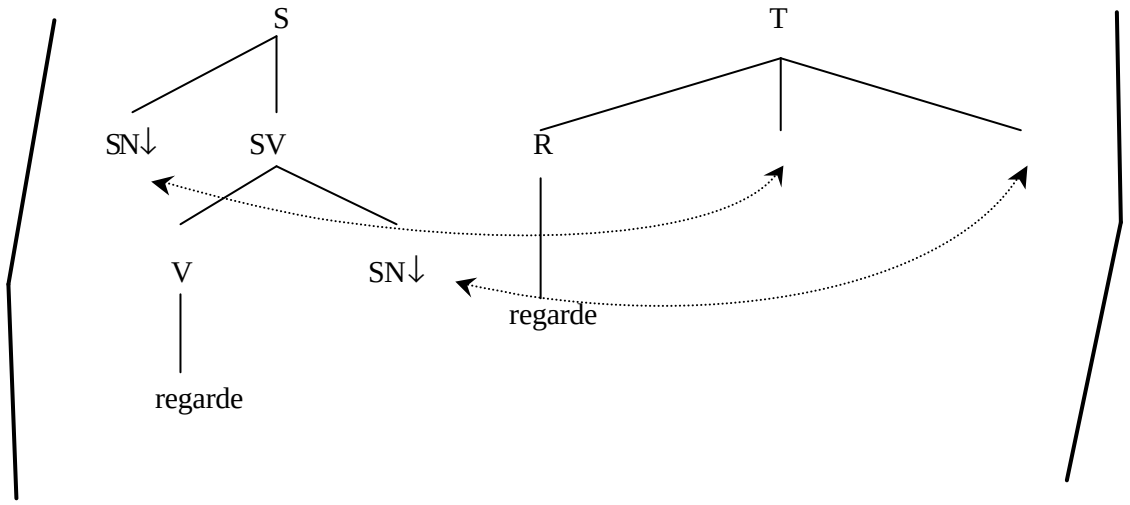


Figure 11. Exemple d'un arbre syntaxique et d'un arbre sémantique synchronisés

L'originalité principale de ce formalisme est que les dérivations syntaxiques et sémantiques doivent être synchronisées. Ainsi, la dérivation de deux arbres $<a_1,a_2>$ se fait selon les étapes suivantes :

1. Choisir de manière non-déterministe un lien entre deux nœuds (n_1 à a_1 et n_2 à a_2)
2. Choisir de manière non-déterministe une paire d'arbres $< b_1, b_2>$de la grammaire.
3. Créer la paire $< b_1 <a_1, n_1>, b_2<a_2, n_2>>$ où $b <a, n>$ est le résultat d'une relation primitive sur a au nœud n en utilisant b.

La traduction automatique est l'application la plus courante de ce formalisme. Le principe de base de ces applications est d'utiliser des règles de transfert d'une langue à une autre. Ainsi, pour chaque arbre de dérivation dans la langue de départ est construit un arbre de dérivation correspondant dans la langue cible. Ceci est fait en établissant un lien entre chaque nœud des deux côtés et en préservant les relations de dominance entre les nœuds dans l'arbre de dérivation source. Le schéma du transfert est présenté dans la figure suivante (Prigent, 1994) :

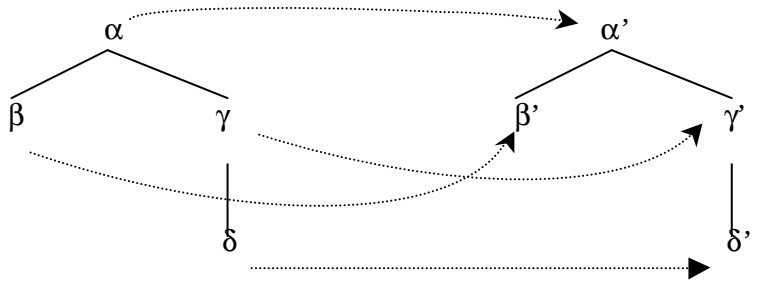


Figure 12. Schéma d'un transfert simple

Une version modifiée de ce formalisme a été proposée pour la traduction automatique de la parole (Cavazza, 1998b).

2.1.1.21 La grammaire d'insertion d'arbres (TIG)

La motivation principale de ce formalisme est de proposer une version du formalisme LTAG équivalente à une CFG et par conséquent analysable en un temps cubique $o(Gn^3)$ (Schabes et Waters, 1995). Ainsi, le formalisme TIG est proposé comme un compromis entre les LTAG et la CFG de manière à combiner l'efficacité computationnelle d'une grammaire CFG au pouvoir expressif d'une grammaire lexicalisée. La TIG est un formalisme basé sur les arbres et qui, tout comme les LTAG, utilise deux opérations : l'adjonction et la substitution. La différence principale entre la LTAG et la TIG est que dans la dernière l'adjonction a été contrainte de manière à éliminer la dépendance au contexte causée par certains types d'adjonction.

2.1.1.22 La grammaire d'arbres furcants (TFG[13])

Le formalisme d'Arbres Furcants TFG s'inscrit dans le contexte des formalismes visant à simplifier le modèle LTAG afin de le rendre plus abordable pour des applications réelles. L'opération de base utilisée dans ce formalisme (la furcation) remonte à (De Smedt et Kempen, 1990), mais la définition générale du formalisme a été faite par (Cavazza, 1998a), et puis développé par (Roussel, 1999). Les différences principales entre la TFG et LTAG peuvent se résumer dans les points suivants :

1. Remplacement de l'adjonction par l'opération de furcation. Cette opération évite l'ajout d'un niveau syntagmatique supplémentaire à la différence de l'adjonction. Ce changement entraîne une simplification syntaxique importante du formalisme et le rend faiblement équivalent à une grammaire indépendante du contexte CFG, à la différence de LTAG qui est un formalisme légèrement dépendant du contexte.
2. Abandon du principe de co-occurrence prédicat argument pour la construction des arbres élémentaires du formalisme.

[13] Tree Furcation Grammar.

3. Adoption du modèle de sémantique interprétative basé sur les travaux de (Rastier, 1987) pour la représentation des traits.

Ce formalisme a fait l'objet d'une implantation dans le contexte d'un système d'analyse robuste de la parole (Roussel, 1999). Cependant, son adaptation au traitement de la parole reste une question ouverte selon les termes de (Roussel, 1999). En effet, le sacrifice de la propriété fondamentale des LTAG (le principe de co-occurrence prédicat argument) au profit d'un ensemble de principe sémantique généraux nous semble un choix discutable. Ces principes sont tellement généraux qu'ils ne sont pas suffisant pour contraindre les arbres et ne permettent pas une intégration efficace des informations supra-linguistiques relatives à la tâche qui sont à la fois fiables et faciles à modéliser.

2.1.1.23 La grammaire stochastique d'arbres adjoints lexicalisés (SLTAG[14])

Les premières versions stochastiques du formalisme LTAG ont été proposées en 1992 par (Resnik, 1992), (Schabes, 1992). Ces modèles sont basés sur les travaux de (Jelinek et al, 19990) sur les SCFGs (les CFGs Stochastiques).

Comme montré dans la figure 13, une CFG stochastique se distingue d'une CFG classique par deux points :

1. Les règles de réécritures sont associées chacune à une probabilité, comme montré dans la figure suivante :

(S_1)	S	→	SN SV	(0.5)
(S_2)	S	→	S SP	(0.35)
(S_3)	S	→	NP VP	(0.15)
(SV_1)	SV	→	V SN	(0.4)
(SV_2)	SV	→	V SP	(0.6)

Figure 13. Fragment d'une grammaire CFG stochastique

2. Calcul d'une probabilité pour chaque dérivation possible. Le calcul de la probabilité d'une est facilité par le fait que chaque réécriture dans une CFG est indépendante du contexte et ainsi la probabilité de la dérivation peut être calculée en multipliant les probabilités des règles de réécritures.

Parallèlement, une SLTAG consiste à associer des probabilités à chaque arbre et puis à sa combinaison avec un autre arbre (par substitution ou adjonction).

Pour une définition formelle de ce modèle, considérons les notations suivantes (Resnik, 1992) :

[14] Stochastic Lexicalized Tree Adjoining Grammar.

- $s(a)$ comme l'ensemble des nœuds frontières de l'arbre a qui sont marqué pour la substitution. Cet ensemble peut être vide dans certains cas.
- $a(a)$ comme l'ensemble des nœuds frontières.
- S(a, a', h) comme la substitution de l'arbre a' et l'arbre a au nœud h.
- A(a, b, h) l'adjonction de l'arbre auxiliaire b et l'arbre a au nœud h et A(a, *non*, h) comme la non-adjonction.
- $\Omega = (s + a)$ l'ensemble des opération de substitution s et d'adjonction a).

Ainsi, une SCFG peut être définie comme un 5-tuple (Resnik, 1992), $<I, A, P_I, P_S, P_A>$ où :

1. I est un ensemble d'arbres initiaux.
2. A est un ensemble d'arbres auxiliaires.
3. P_I est une fonction de I dans l'intervalle [0,1], tel que $\Sigma P_I(\alpha) = 1$. Cette fonction représente la probabilité qu'une dérivation soit à partir de l'arbre α.
4. P_S est une fonction de Ω dans l'intervalle [0,1] tel que $\forall a \in I \cup A, \forall h \in s(a) \Sigma / a \hat{I} I /$ Ps(S(a, a', h) = 1.
5. P_A est aussi une fonction de Ω dans l'intervalle [0,1] tel que $\forall a \in I \cup A, \forall h \in a(a) \Sigma / b \hat{I} A$ *none* I / P_A (A(a,b', h) = 1.

Ainsi, le formalisme SLTAG présente trois avantages principaux comme cadre pour l'analyse des langues naturelles (Resnik, 1992), (Joshi et Shabes, 1999) :

1. Le principe de co-occurrence prédicat argument évite les problèmes liés à la taille de la fenêtre dans les approches à base de n-grammes. Ainsi, on associe une seule probabilité à tous les éléments liés syntaxiquement plutôt que d'associer à chacun une probabilité différente.
2. LTAG étant un formalisme lexicalisé, les probabilités associées aux opérations structurales sont aussi sensibles au contexte lexical. Cette prise en considération du contexte lexical n'est pas faite au détriment de l'indépendance des probabilités des opérations puisque les adjonctions et les substitutions dans des nœuds différents sont indépendants les uns des autres.
3. Représentation flexible du lexique permettant de représenter les arbres ancrés par un ou plusieurs mots ainsi que les schèmes syntagmatiques, ce qui permet une représentation économique pour le traitement de certains cas comme les expressions idiomatiques.

Malgré ces avantages, le formalisme SLTAG présente des inconvénients pratiques notamment en ce qui concerne la taille importante des données nécessaires à l'apprentissage des paramètres de la grammaire.

2.2 La grammaire sémantique

Depuis le début de la philosophie, notamment avec Aristote, les philosophes du langage ont distingué entre trois éléments : la parole, les états de l'âme et les choses. Cette distinction triadique s'est cristallisée plus tard avec les philosophes du moyen âge comme saint Thomas d'Acquint qui la reformule ainsi (cité dans (Rastier, 1991, page 75)) : *les paroles sont les signes des pensées et les pensées des similitudes des choses*. Ce qui signifie que, selon cette distinction, les paroles se réfèrent aux choses moyennant les concepts. La triade de la signification est présentée dans la figure suivante.

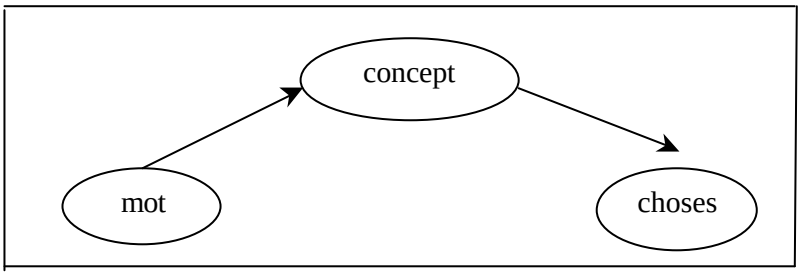


Figure 14. La triade de la signification

La distinction entre structure conceptuelle et structure linguistique a trouvé un regain d'intérêt dans la linguistique moderne avec les grammaires générales, les travaux de ((Ogden et Richards, 1923), cité dans (Rastier, 1991)) sur la sémantique et elle continue à être une idée fondamentale dans les travaux actuels en philosophie du langage.

Par ailleurs, plusieurs travaux dans les domaines de la psychologie et de l'intelligence artificielle, ont montré que la mémoire humaine est organisée selon des schémas qui organise les concepts suivant leurs relations sémantiques et leur pertinence par rapport au contexte (Minsky, 1975), (Kuipers, 1975), (Schank, 1977), (Schank et Abelson, 1977).

C'est dans ce contexte qu'est née la grammaire sémantique (Burton, 1975) qui est, à notre connaissance, le premier à proposer la grammaire sémantique.

De nos jours, différentes formes de ce formalisme ont été proposées. Ces formes sont tellement différentes qu'elles éclipsent les points communs et amènent même à penser qu'il ne s'agit pas du même formalisme. Cette différence est accentuée par la rareté des études linguistiques de ce formalisme[15]. Par exemple, au niveau terminologique, certains chercheurs parlent de grammaire sémantique (Burton,1975), (Ward, 1991), (Gavaldà, 2000), grammaire conceptuelle (Perennou, 1996), sémantique globale (De Mori, 1994), grammaire de cas (Minker *et al.*, 1996).

Dans ce qui suit, nous allons nous efforcer de présenter ce formalisme dans ses aspects généraux indépendamment des différences superficielle et terminologique.

[15] A notre connaissance, la seule exception est le travail de (Perennou, 1996).

2.2.1 Les bases linguistiques de la grammaire sémantique

Comme nous avons vu dans les paragraphes précédents, les formalismes syntaxiques comme LTAG permettent d'analyser les énoncés en procédant de manière sémasiologique, c'est-à-dire en commençant par la nature morpho-syntaxique de chaque mot, en examinant son environnement syntaxique et puis en lui associant une structure sémantique et puis pragmatique.

Contrairement aux formalismes classiques, la grammaire sémantique procède de manière onomasiologique, c'est-à-dire le point de départ de la grammaire est la représentation pragmatique finale (appelée sémantique par abus de langage) et puis dans une deuxième étape l'association à cette structure *idéalement* de toutes les formes possibles de sa réalisation dans le contexte d'utilisation du système.

Afin de mettre au clair cette différence importante, examinons l'exemple suivant :

J'arrive mardi prochain (40)

L'analyse de cet énoncé selon l'approche sémasiologique se fait selon les étapes suivantes :

- **Analyse syntaxique :** cette phase consiste à associer un arbre syntaxique à l'énoncé selon les principes que nous avons vu dans le paragraphe précédent sur le formalisme LTAG.
- **Analyse sémantique :** association des rôles aux mots : agent, patient, etc.
- **Analyse pragmatique et discursive :** association des fonctions pragmatiques, thème, rhème, etc.

Dans le contexte d'une approche onomasiologique comme la grammaire sémantique, le point de départ est l'établissement d'une liste contenant l'ensemble des unités sémantiques susceptibles d'être utilisées dans le contexte d'une application. Différents critères ont été proposés dans la littérature pour l'établissement de cette liste en particulier pour définir les unités sémantiques. Malgré leurs divergences, ces critères ont un commun un pragmatisme qui les pousse à choisir les unités qui facilitent le plus l'implantation, sans se soucier des aspects linguistiques théoriques.

Par exemple, dans les énoncés suivants :

1 **Je voudrais** réserver une chambre pour deux personnes.

2 **C'est pour** réserver une chambre pour deux personnes.

3 **Je vous appelle pour** réserver une chambre pour deux personnes.

On remarque que ces trois énoncés ont le même sens (dans le contexte d'un dialogue de réservation hôtelière). Ainsi, nous pouvons considérer que les segments : *je voudrais, c'est pour* et *je vous appelle pour* ont la même fonction et peuvent donc être associés à la même catégorie sémantique [formule_demande].

Les postulats de base derrière la grammaire sémantique sont les suivants :

- L'unité sémantique est exprimée dans un énoncé par plusieurs mots ou une séquence de mots (qui apparaissent dans l'énoncé).

- Le sens général de l'énoncé peut être représenté par l'ensemble des unités sémantiques.

D'un point de vue formel, la grammaire sémantique est équivalente à une grammaire de type CFG (Gavaldà, 2000) mais, à notre connaissance, aucune vraie définition formelle n'a été donnée de ce formalisme; les seules indications données à ce propos sont généralement que la grammaire sémantique est une CFG dans laquelle les non-terminaux peuvent être de nature sémantique.

Pour rendre les idées de ce formalisme examinons l'exemple suivant de grammaire sémantique :

[my_unavailability][16]

(i *BABBLE CANT *MEET +[temporal]

(+[temporal] BE *BABBLE BAD *FOR_ME)

BABBLE		BE	
	(really)		(is)
	(probably)		(would be)
	(kind of)	BAD	
	(unfortunately)		(bad)
CANT			(tight)
	(can't)		(booked solid)
	(Couldn't)		(Packed)
	(don't want to)		(out)
MEET			(no good)
	(Meet)	FOR ME	
	(do it)		(for me)
	(make it)		(here)

Figure 15. Un exemple de grammaire sémantique classique (Mayfield *et al.*, 1995)

La première remarque que nous pouvons faire à propos de cette grammaire, c'est que la catégorisation des mots est faite sur des critères purement sémantiques : on ne distingue pas si un mot est, par

[16] Les mots marqués avec * sont facultatifs, les mots marqués avec + sont des mots qui peuvent se répéter. Les mots en lettres capitales sont des non-terminaux dont les réécritures sont présentées entre parenthèses. Les mots entre crochets correspondent à des expressions spéciales.

exemple, un adverbe ou un nom. Deuxièmement, on peut voir que les unités sont, elles aussi, définies sur des critères purement sémantiques et que la syntaxe bien que présente au sein de certains de ces segments (par l'intermédiaire de l'ordre des mots) n'est pas exprimée de manière déclarative.

Par exemple, le non-terminal (CANT) est représenté comme l'ensemble des formes de verbes négatifs. Cet amalgame entre la syntaxe au sein de la sémantique nous oblige à recréer une deuxième règle pour le non-terminal *CAN* (la forme affirmative du verbe). Donc, à chaque fois que nous allons stoker un non-terminal correspondant à une structure dans laquelle il y a un verbe affirmatif nous sommes obligés de créer une structure négative équivalente et l'associer à un autre non-terminal. En d'autres termes, pour exprimer la négation dans un formalisme syntaxique, il faut un nombre très limité de règles alors que dans la grammaire sémantique il faut N règles où N est le nombre des constructions verbales dans la grammaire.

2.2.2 Portée et limites de la grammaire sémantique

Ce formalisme est le plus couramment utilisé dans le contexte des systèmes de traitement automatique la parole. Ceci est dû à une série d'avantages :

1. **Sur le plan computationnel :** équivalence forte avec les grammaires CFG, pour lesquelles il existe plusieurs algorithmes dont le temps d'analyse est cubique.
2. **Sur le plan de la tâche d'analyse :** ce formalisme permet d'augmenter la robustesse de l'analyse, étant donné que la dimension syntaxique y est limitée à l'ordre des mots implicitement. Ainsi, le système évite la plupart des erreurs syntaxiques qui peuvent résulter de problèmes de reconnaissances ou autres.
3. **Sur le plan pratique :** la mise en œuvre de ce formalisme est plus facile que pour les grammaires syntaxiques classiques et nécessite moins d'expertise en linguistique.

Malgré ces avantages, la grammaire sémantique ne constitue pas une solution idéale pour le traitement des dialogues dans les domaines limités et encore moins pour le traitement de textes ouverts. Les principaux inconvénients de ce formalisme se résument dans les points suivants :

1. La grammaire sémantique n'a pas un statut linguistique et formel bien défini. Cela réduit les possibilités de comparaison objectives avec les autres formalismes et rend son choix pour une application quelconque une tâche difficile. Cela rend aussi la tâche de l'enseignement de ce formalisme plus difficile.
2. L'interaction directe entre les connaissances linguistiques et l'univers conceptuel de la tâche, qui est l'avantage principal de la grammaire sémantique, est aussi son principal inconvénient. En effet, la dépendance de la tâche réduit considérablement la portabilité de la grammaire vers d'autres domaines applicatifs et rend obligatoire l'écriture d'une nouvelle grammaire à chaque changement de domaine.

3. Ce formalisme est adapté à des applications de petite taille généralement et présente des difficultés pour des applications dont le domaine est large en particulier pour représenter les relations sémantiques entre les différentes unités (Pieraccini et Levin, 1995).
4. A cause de la réduction du rôle de la syntaxe dans la grammaire sémantique ne permet pas de refléter facilement certaines nuances sémantiques exprimées par des phénomènes syntaxiques complexes. En effet, l'expression des contraintes syntaxiques n'est pas une procédure économique en terme de nombre de règles comme c'est le cas de la négation.

2.2.3 Extensions de la grammaire sémantique

A notre connaissance, il n'y a pas eu de vraies extensions de la grammaire sémantique. Cependant, des aménagements ont été faits de ce formalisme afin de l'adapter au traitement de dialogues dont le domaine est large (les dialogue dits multi-domaine). L'un des principaux travaux dans ce contexte, est la Grammaire Sémantique Modulaire proposée par les chercheurs du ISL-CMU (Woszczyna *et al*, 1998). Comme l'indique son nom, ce formalisme permet de séparer les grammaires des sous-domaines en fichiers indépendants qui se complètent de manière modulaire. Les auteurs de ce formalisme énumèrent les avantages suivants (Wosczyna *et al*, 1998) :

1. La séparation des grammaires des sous-domaines permet à différents linguistes de travailler en parallèle pour l'écriture de la grammaire sans interférence de leurs travaux.
2. Création d'une grammaire inter-domaine (qui contient des expressions de temps, date, politesse, etc.) dont l'utilité est le maintien de la consistance de l'analyse et l'augmentation de la portabilité du système, étant donné que cette grammaire peut être utilisée dans un bon nombre d'applications.
3. D'un point de vue ingénierie du logiciel, la séparation des sous-grammaires permet de distinguer le domaine correspondant à chaque énoncé. La reconnaissance du domaine de l'énoncé permet de résoudre certaines ambiguïtés causées par l'élargissement du domaine.

Malgré les avantages de cette modularité, la dépendance à la tâche reste une limitation importante de ce formalisme. De plus, les problèmes liés au traitement des phénomènes syntaxiques et à la finesse de l'analyse avec la grammaire sémantique classique restent complètement posés avec ce formalisme.

3 Chapitre I.3 : Les approches d'analyse robuste du langage oral

Afin de prendre en considérations les différentes sources de manque de robustesse à l'oral, différentes techniques ont été proposées et testées dans la littérature dans des contextes applicatifs divers. Nous distinguons entre deux types d'approches : les approches pour l'analyse syntaxique robuste et les approches pour le traitement des extragrammaticalités.

3.1 Les approches pour l'analyse syntaxique robuste

Comme nous l'avons dit au début de cette thèse, un système robuste est un système qui est capable de fournir une analyse correcte même dans les cas d'une entrée déformée ou inattendue. Les différentes techniques d'analyse robuste ont été développées dans le cadre de travaux sur l'oral tout comme sur l'écrit. Dans les deux cas, l'objectif des travaux est l'utilisation des algorithmes d'analyse dans des conditions réelles : erreurs de reconnaissance et extragrammaticalités pour la parole et fautes de frappe et erreurs grammaticales dans les textes écrits. Les principales techniques utilisées dans l'analyse robuste consistent en des extensions d'algorithmes classiques d'analyse afin de les dynamiser et les rendre plus adaptés aux inattendus des applications réelles. Par ailleurs, certaines approches se sont inspirées de travaux dans des domaines relativement loin comme la recherche d'informations et la classification de documents.

3.1.1 L'analyse partielle par segments (chunking)

2.1.1.24 Principes généraux

Inspiré par les travaux de (Gee et Grosjean, 1983) en psycholinguistique, (Abney, 1991), (Abney, 1995) propose une approche d'analyse partielle basée sur le segment (chunk parsing). Les segments, considérés comme unité de base de traitement, sont des structures syntaxiques correspondants à un graphe connecté dans l'arbre d'analyse d'un énoncé. Ces unités sont définies selon leurs *têtes syntaxiques majeures*. Les têtes syntaxiques sont généralement des *mots à contenu* (non grammaticaux) à l'exception des cas où un mot apparaît entre un mot grammatical *mg* et un mot à contenu que sélectionne *mg*.

Dans un système d'analyse par segment, le processus d'analyse est divisé en deux parties complètement distinctes (contrairement aux approches classiques dans lesquelles les deux étapes sont fusionnées) :

- **La segmentation :** il s'agit de convertir le flux de mots en un flux de segments.

- **L'attachement :** consiste à attacher les segments obtenus dans la phase précédente au sein d'une structure globale qui set l'arbre d'analyse de l'énoncé. Concernant à la partie précédente, cette étape n'est pas obligatoire ou au moins elle n'est pas systématique. Ainsi, un analyseur partiel peut fournir des arbres d'analyse complets et des segments partiels ou des segments partiels uniquement.

Différentes approches similaires à celle d'Abney ont été proposées, comme celle de (Aït-Mokhtar et Chanod, 1997), (Grefenstette, 1999) basée sur des techniques de FSAs et celle du supertagging proposée par (Srinivas, 1996), (Srinivas, 1997), dans le cadre du formalisme LTAG.

2.1.1.25 Le système CASS

CASS (Cascaded Analysis of Syntaxctic Structure) est un système d'analyse syntaxique robuste à base de segments. Ce système a été développé par Steven Abney à l'université de Tübingen en Allemagne (Abney, 1991), (Abney, 1996). CASS utilise un ensemble d'analyseurs simples qui s'appliquent en cascade pour construire une représentation syntaxique globale de l'énoncé.

L'entrée de CASS est la sorite du module d'analyse morphologique de Church qui fournit les POS tags aux mots ainsi que les syntagmes nominaux simples (non-récursifs). Notons que le taux de traitement des syntagmes nominaux est inférieur à celui des POS tags. Le traitement de cette entrée dans le système se fait selon trois étapes :

3.1.1.1.1 Le filtre des segments

Ce module est basé sur deux sous-filtres :

1. Le filtre des syntagmes nominaux : ce module utilise des expressions régulières pour assembler les syntagmes nominaux sur la base de l'analyse superficielle fournie par le reconnaisseur de syntagmes nominaux de Church. De même ce module corrige les erreurs de traitement des syntagmes nominaux par le module de Church comme ceux résultants des adjectifs prénominaux.
2. Le filtre des segments : ce module utilise aussi des expressions régulières pour reconnaître le reste des segments. Voici un exemple de la sortie de ce module avec l'énoncé : *In south Australia beds of boulders were deposited*.

CS

[pp in [Np south Australia beds]]

[pp of[Np boulders]]

[Vp were deposited]

.CS

Comme nous pouvons le voir le système a commis une erreur d'analyse (à cause du tagger) du premier syntagme nominal *south Australia beds*.

3.1.1.1.2 Le filtre des propositions

Le filtre des propositions consiste en deux sous-filtres :

1. Le filtre brut : ce filtre essaie de reconnaître les frontières des propositions simples ainsi que de marquer le sujet et le prédicat de la proposition. S'il n'arrive pas à identifier un seul sujet ou prédicat, ce module identifie le type d'erreur rencontré comme l'existence de plusieurs syntagmes verbaux ou l'absence du sujet (à cause d'une ellipse par exemple), etc.
2. Le filtre des propositions corrigées : ce module essaie de corriger les erreurs identifiées par le module précédent en appliquant des patrons spécifiques à chaque cas. Voici par exemple le patron utilisé pour la correction des complémenteurs non-analysés : [pp X_p-time NP] … VP $\rightarrow$ [$_{clause}$ X_c NP ... VP]. Au cas où aucun des patrons n'est pas applicable à l'entrée, le système utilise des heuristiques générales qui lui permettent d'améliorer l'analyse sur la base d'informations partielles (comme l'existence d'une syntagme nominal à coté d'un syntagme verbal, un syntagme verbal seul, etc.). Ainsi, après cette étape, l'analyse obtenue pour l'énoncé devient comme suivant :

   [$_{pp}$ in south Australia]

   [$_{Subj}$ [Np beds]]

   [$_{pp}$ of boulders]

   [$_{Pred}$ [Vp were deposited]]

   Comme nous pouvons le voir dans l'analyse précédente, le système a réussi à corriger l'erreur d'analyse dans le premier sntagme.

3.1.1.1.3 Le filtre d'analyse

Contrairement aux modules précédents, le filtre d'analyse est basé sur des règles récursives (pas des expressions régulières). La fonction principale de module est d'assembler les structures récursives en attachant les nœuds les uns aux autres selon la nature des têtes de ces structures et les contraintes grammaticales sur leur assemblage. Par exemple, un segment *Y* peut être attaché à un segment *X* seulement si la tête de *X* peut avoir *Y* comme argument ou modifieur.

Les résultats de Cass ont montré qu'il est à la fois assez robuste et très rapide pour le traitement des corpus écrits. Ces résultats sont principalement dus à l'architecture de ce système qui consiste à appliquer différents niveaux d'analyse en cascade avec des règles et des patrons qui permettent de corriger les erreurs effectuées dans les étapes précédentes.

3.1.2 Les approches sélectives

2.1.1.26 Principes généraux

Les approches sélectives consistent à n'analyser que les parties jugées pertinentes de l'énoncé reçu. Ces approches sont appuyées par des observations simples sur le traitement humain de la parole qui est caractérisé par la variation du degré de l'attention. D'un point de vue informatique, il s'agit souvent

d'équiper l'algorithme d'analyse par un filtre qui permet, selon un certain nombre de contraintes, d'ignorer un(des) mot(s) ou les segments non pertinents ou non analysables.

Différents degrés de sélectivité ont été utilisés dans la littérature. Cela varie entre des approches assez proches de l'analyse à base de mots clés comme (Luzzati, 1987), (Rouillard, 2000) jusqu'à des approches à couverture plus raisonnable comme l'algorithme GLR* de (Lavie, 1997) ou les différentes implantations des grammaires sémantiques à Carnegie Mellon University, (Mayfield, 1995), (Gavaldà, 2000), (Bousquet, 2002). Contrairement à ce que certains chercheurs dans le domaine pensent, les approches sélectives ne sont pas forcément synonymes de perte d'information ou d'analyse superficielle. En effet, une stratégie sélective bien conçue peut être ajoutée à n'importe quel système d'analyse syntaxique sans affecter sa profondeur d'analyse. Le seul inconvénient de ces approches est qu'elles augmentent la complexité computationnelle des algorithmes auxquels elle est ajoutée. Par exemple, (Wang, 2001) décrit un algorithme de type *chart* augmenté par une stratégie sélective (pour l'analyse d'une grammaire sémantique équivalent à CFG) dont la complexité est $O(n^4)$ au lieu de $O(n^3)$ comme c'est le cas de plusieurs algorithmes classiques pour la CFG[17].

2.1.1.27 Le système Phoenix

Les chercheurs de l'ISL-CMU (Interactive Systems Labs. à Carnegie Mellon University) ont adopté une approche à base de grammaires sémantiques stochastiques pour leur système ATIS. L'approche adoptée est basée sur un module d'analyse syntactico-sémantique qui a pour entrée la sortie du système de reconnaissance et dont la sortie est traitée par un module d'analyse sémantique.

L'architecture générale de ce système est présentée dans le schéma suivant :

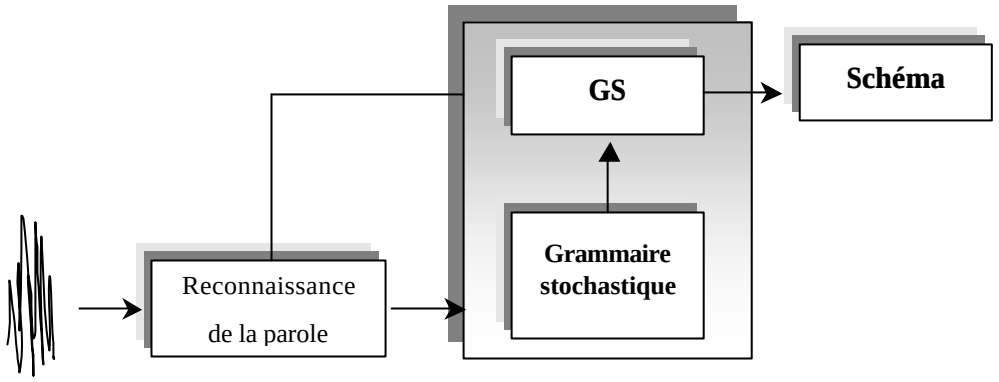


Figure 16. Architecture du système ATIS du ISL-CMU

Comme nous pouvons le voir dans la figure précédente, l'analyse syntactico-sémantique se fait en deux étapes :

[17] Cette information est mentionnée indirectement dans l'article de Wang mais elle a été donnée explicitement au cours de l'exposé oral de cet article à la conférence Eurospeech 2001 à Aalborg au Danemark.

1. **Analyse superficielle de la parole :** le treillis de mots (sortie du module de reconnaissance) est analysé tout d'abord par une grammaire stochastique (de paires) relativement lâche, afin de pouvoir tolérer les parties qui contiennent des extragrammaticalités.
2. **Le module d'analyse à base de Grammaire Sémantique :** ce module a deux tâches principales. Tout d'abord, il essaye d'analyser les parties extragrammaticales (tolérées par la grammaire de bi-grammes) en lui imposant des contraintes supplémentaires à l'aide d'une grammaire sémantique. Ensuite, il traduit la représentation sémantique de l'énoncé en schéma. La grammaire sémantique utilisée est convertie en un RTR stochastique capable de résoudre les ambiguïtés conceptuelles. A titre d'exemple, la règle présentée dans la figure 15 a été implémentée sous la forme du réseau de transition suivant :

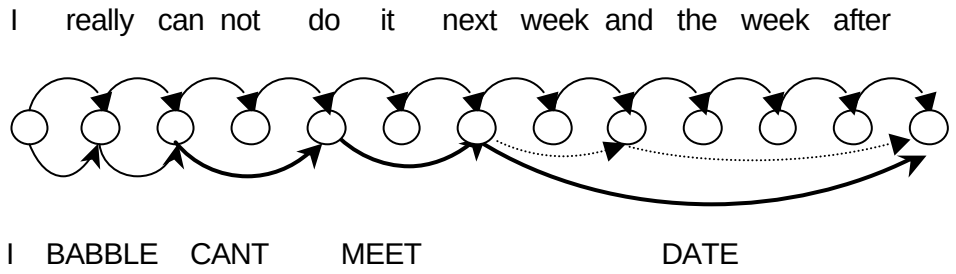


Figure 17. Exemple de réseau de transition récursif utilisé par phœnix

Les réseaux implantés n'ont pas de contraintes particulières sur le nombre des slots à reconnaître. Ainsi, l'analyseur est capable de traiter tous les concepts pertinents qui se trouvent dans la chaîne d'entrée. Il est aussi capable d'ignorer tous les mots qui ne font pas partie du lexique de l'analyseur et qui figurent entre ces concepts (et pas à l'intérieur d'un segment conceptuel), mais il est par contre, incapable de traiter les mots du lexique qui figurent dans des positions non pertinentes. Ces dernières conduisent l'analyseur à ignorer tout simplement le concept en cours d'analyse (on échoue à le reconnaître) mais ne conduisent pas à l'échec total de l'analyse des autres concepts (Mayfield *et al.*, 1995).

3.2 Les approches pour le traitement des extragrammaticalités de l'oral

3.2.1 Introduction

Comme nous avons vu dans la première partie, les extragrammaticalités de l'oral jouent un rôle important dans le traitement des dialogues. A cause de cette importance, ce phénomène a fait l'objet de nombreuses études. A notre connaissance, les premières études des extragrammaticalités remontent au début du siècle passé avec les travaux en psychanalyse menés notamment par (Freud, 1901) et dont l'objectif était d'analyser l'intention cachée du locuteur en observant ses lapsus. Au-delà de ces

premiers travaux, les extragrammaticalités de l'oral ont fait l'objet d'études descriptives et applicatives dans des disciplines diverses qui dépasse largement notre centre d'intérêt ici :

1. **Etudes linguistiques :** différentes études descriptives ont été menées afin de caractériser les principaux aspects linguistiques des extragrammaticalités en particulier leurs propriétés acoustiques et prosodiques. A titre d'exemple, nous pouvons citer les travaux de (Hockett, 1967) et l'analyse de corpus de (Nakatani et Hirshberg, 1994).
2. **Etudes psychologiques :** la partie principale de ces études porte sur la production des extragrammaticalités par les humains selon plusieurs points de vue comme les différences individuelles de production et l'unité psycholinguistique de traitement (Maclay et Osgood, 1967), l'interaction entre la structure des extragrammaticalités et leurs processus de génération (Levelt, 1983), le rôle de facteurs communicatifs (connaissance des locuteurs, familiarité avec le thème de la conversation, etc.), (Fox Tree et Schrock, 1999), la production des extragrammaticalités par les aphasiques (Hartsuiker et Kolk,1998), etc. Des études moins nombreuses ont porté sur la perception des extragrammaticalités. Nous pouvons citer sur ce sujet les travaux de Robin Lickley à l'université d'Edinburgh (Lickley, 1994).
3. **Etudes applicatives en traitement automatique de la parole :** plusieurs travaux ont porté sur la détection et la correction des extragrammaticalités de la parle spontanée. Parmi les premiers travaux, nous pouvons citer (Carbonell et Hayes, 1983) qui ont proposé l'utilisation de patrons simples pour traiter certaines extragrammaticalités. Dans la même période (Hindle, 1983) propose une approche syntaxique déterministe pour la correction des patrons (il considère le point d'interruption comme étant déjà détecté). Par ailleurs**,** des travaux récents ont porté sur la production des extragrammaticalités dans le cadre d'un moteur de génération incrémental (Finkler, 1997).

Dans les paragraphes suivants, nous allons nous limiter aux approches relativement récentes dans le domaine du traitement automatique de la parole.

3.2.2 L'approche « *d'analyse d'abord »* de SRI international

Fondé sur 607 énoncés contenant des extragrammaticalités extraits du corpus ATIS (Air Travel Information Services), le travail de (Bear *et al.*, 1992), (Shriberg, 1994) est l'un des premiers à reprendre les travaux sur les extragrammaticalités dans un cadre applicatif.

2.1.1.28 Le schème d'annotation

La première étape de ce travail consistait à proposer un schème de notation qui combine la simplicité à la finesse nécessaire pour la représentation des différentes formes d'extragrammaticalités. Les aspects de base de ce schème de notation sont les suivants (Bear *et al.*, 1992) :

1. Le point d'interruption est représenté par une barre verticale (|).

2. Correspondance identique : pour montrer que deux mots aux deux côtés d'une interruption sont identiques, on les marque *M* (*M* est la première lettre du mot anglais *matching*).
3. Le remplacement : indique le remplacement d'un mot avant le point d'interruption par un mot après. Les deux mots doivent être similaires morphologiquement. En général ils doivent être de la même catégorie ou d'une variante morphologique de celle-ci comme les cas d'amalgames : I/I'd.
4. Mots neutres : tous les mots dans la zone d'une extragrammaticalité est noté *X*.
5. Un tiret (-) est ajouté aux signes précédents en cas d'incomplétude.

Voici quelques exemples de la notation.

I want fl- flights to Boston
M_1- | M_1

What what are the fares
M_1 | M_1

Show me flights daily flights
M_1 | X M_1

2.1.1.29 La détection et correction des extragrammaticalités

L'approche proposée consiste à combiner deux techniques :

1. **Analyse syntaxique et sémantique :** Afin de réduire les surgénérations des patrons, les chercheurs de SRI ont utilisé les modules d'analyse syntaxique et sémantique du système GEMINI qui est une re-implantation du *core language engine* (Alshawi, 1992).
2. **Reconnaissance de patron (pattern matching) :** cette technique est utilisée pour détecter les phénomènes simples tel que, la répétition d'une séquence de mots comme *I would like a book I would like a flight* ou des anomalies syntaxiques simples comme : «*a the*», ou «*to from*», etc.

Ainsi, l'analyse se fait selon deux étapes : tout d'abord le système tente d'analyser les énoncés syntaxiquement et sémantiquement et puis dans la deuxième phase, il passe les énoncés au *reconnaisseur de patrons*. Dans ce cas, deux types de décision sont possibles :

- Les parties d'énoncés qui ont été correctement traitées par les modules d'analyse syntaxique et sémantique et qui sont signalées comme étant extragrammaticales par le reconnaisseur de patrons sont considérés comme des surgénérations (false-positive cases).
- Les parties d'énoncés incomplètement analysées par les modules linguistiques et qui sont signalées par le reconnaisseur de patrons comme étant extragrammaticales sont considérées comme étant des extragrammaticalités réelles.

L'inconvénient principal de cette combinaison est qu'elle est incompatible avec les approches d'analyse partielle qui sont les plus adaptées au traitement de l'oral. Cela nous met devant un dilemme :

D'une part, l'utilisation d'une méthode d'analyse partielle (qui réussit pratiquement toujours à donner une analyse) nous empêche de juger la grammaticalité d'un énoncé et par conséquent rend ce type de combinaison impossible. D'autre part, les méthodes d'analyse classiques sont bien adaptées pour le jugement de grammaticalité (tous les énoncés analysés sont complètement corrects grammaticalement) mais elles échouent souvent à traiter correctement des phénomènes syntaxiques propres ou fréquents à l'oral comme les problèmes d'accord, les ellipses, etc. Par ailleurs, des échecs causés par l'un de ces phénomènes peut conduire à une erreur de jugement d'une extragrammaticalité. De plus, le jugement de non-grammaticalité d'un énoncé n'est pas informatif concernant la surgénération d'un patron lorsqu'on a un énoncé avec plusieurs segments détectés comme correspondant à des extragrammaticalités : on ne sait pas si tous les segments sont réellement extragrammaticaux ou si seulement certains d'entre eux le sont. Finalement, cette approche rend le module de traitement des extragrammaticalités complètement dépendant de l'analyseur syntaxique et par conséquent elle réduit considérablement sa portabilité (on ne peut pas utiliser le module de traitement des extragrammaticalités avec d'autres systèmes).

Les résultats obtenus par (Bear *et al.*, 1992) pour la correction des extragrammaticalités sont 43% de rappel et 50% de précision. (Dowding *et al.*, 1993) a utilisé les mêmes données d'apprentissage avec des modifications légères sur l'entraînement a obtenu un rappel de 30% et une précision de 62%.

3.2.3 L'approche stochastique à basede patrons de Heeman

Ce travail est réalisé dans le cadre du projet américain *TRAINS* à l'université de Rochester. Le corpus utilisé a été spécialement collecté par (Heeman et Allen, 1995) pour étudier les extragrammaticalités de l'oral[18].

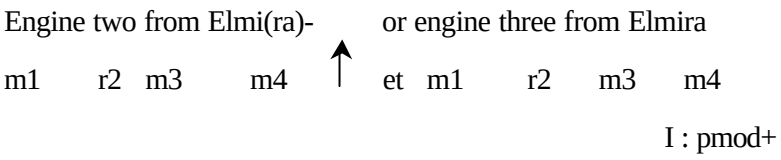
2.1.1.30 Le schème d'annotation

La première étape du travail de Heeman a consisté à proposer une version modifiée du schème d'annotation des chercheurs de SRI. Les principaux symboles utilisés dans ce schème sont : *ipr* pour marquer le point d'interruption. Une série de suffixe est utilisée pour marquer le type d'extragrammaticalité comme : *mod* pour les patrons *modification repairs*, *can* pour les faux-départs *cancels*, et pour les mots d'édition *editing terms*. Les cas ambigus sont marqués par un (+) à la fin. La différence principale entre le schème de Heeman et celui de SRI est que celui de Heeman ne permet pas le partage de la zone remplacée dans le cas d'extragrammaticalités imbriquées.

L'annotation concerne les répétitions, les patrons et les faux-départs. Tous les cas qui couvrent une partie d'un mot ou plus ont été considérés dans l'analyse.

Voilà un exemple d'un cas annoté selon le schème de Heeman :

[18] Une présentation détaillée de ce corpus sera faite dans la troisième partie de cette thèse.



Enoncé (d93-15.2 utt42)

Figure 18. Un exemple d'une extragrammaticalité annotée selon le schème de (Heeman, 1997)

2.1.1.31 La méthode de détection et de correction des extragrammaticalités

Différentes sources d'informations ont été utilisées dans la détection et la correction des extragrammaticalités. Ces sources couvrent l'identité des mots (pour les répétitions), des informations syntaxiques de bas niveau, les transitions entre les mots et les indices acoustiques et prosodiques (en particulier le silence). Suite à l'annotation des extragrammaticalités, Heeman obtient 1302 cas d'extragrammaticalités avec 160 structures différentes (Heeman, 1997). Afin d'éviter les surgénérations de certains patrons, Heeman propose une série de règles pour les contraindre (Heeman et Allen, 1994). Ces règles portent essentiellement sur la forme de la zone d'édition et sa localisation par rapport au point d'interruption d'une part et le reste de l'extragrammaticalité d'autre part. Par ailleurs, pour intégrer les différentes sources de connaissance, il utilise un modèle d langage basé sur les catégories morpho-syntaxiques plutôt que sur les mots.

Ainsi, il utilise un modèle de langage dans lequel plusieurs variables (correspondant aux différentes sources de connaissances) sont utilisées :

$$W'P'R'E'T'S' = \arg\max_{WPRET} \Pr(WPRETS|A) \quad (1)$$

$$= \arg\max_{WPRET} \frac{\Pr(A|WPRETS)\Pr(WPRETS)}{\Pr(A)} \quad (2)$$

$$= \arg\max_{WPRET} \Pr(A|WPRETS)\Pr(WPRETS) \quad (3)$$

Où *W'* est la séquence de mots d'entrée, *P'* la séquence des étiquettes morphologiques (POS tags) correspondant à *W'*, *R'* est l'ensemble des variables d'une extragrammaticalité (Repair), *E'*est l'ensemble des mots d'une zone d'édition, *T'* correspond aux tons, S' au silence et *A* au signal de parole. Dans l'équation (3), le premier terme correspond au modèle acoustique et le second correspond au modèle de langage. Ainsi, le modèle de langage peut être représenté comme suivant :

$$\Pr(W_{1,N}P_{1,N}R_{1,N}E_{1,N}T_1, {}_NT_{1,N}) \quad (4)$$

Où *N* est le nombre de mots dans la séquence d'entrée.

Le silence ainsi que les fragments de mots (considérés comme une partie de la zone d'édition) sont aussi utilisés dans le processus de traitement. Ces indices sont certes importants dans la détection d'une extragrammaticalité, mais le problème est que ces sources d'information ne sont pas fiables

avec une sortie de reconnaissance réelle : les mots incomplets n'étant pas reproduits par les systèmes de reconnaissance et les silences n'étant pas faciles à détecter par les modèles acoustiques.

Dans leur article (Heeman & Allen, 1994) présentent un cas d'extragrammaticalité imbriqué et montrent très sommairement comment leur système le traite sans donner aucune information sur le mécanisme de contrôle qui est le point clé dans ce genre de situations. Etant donné qu'un traitement avec un algorithme gauche-droite classique est incapable de prendre en considération ce phénomène puisqu'il est incapable d'assigner deux catégories différentes à un même mot, nous avons déduit de cet exemple que le système réinitialise le traitement à chaque détection et correction d'une extragrammaticalité. Ainsi, Le système analyse l'énoncé une seule fois au cas de non-existence d'extragrammaticalité et dans le cas d'occurrence d'extragrammaticalités il l'analyse *N+1* fois où *N* est le nombre des occurrences des cas d'extragrammaticalité, ce qui ne nous semble pas être une solution économique.

Les résultats obtenus par Heeman sont présentés dans le tableau suivant :

Phénomène	**Action**	**Rappel**	**Précision**
Discontinuités	Détection	75.88	82.51
	Correction	75.65	82.26
Réparations	Détection	80.87	83.37
	Correction	77.95	80.36
Faux-départs	Détection	48.58	69.21
	Correction	36.21	51.59
Total	Détection	76.79	86.66
	Correction	65.85	74.32

Tableau 1. Les résultats obtenus par (Heeman, 1997) sur la détection et la correction des extragrammaticalités

Comparés aux résultats obtenus par (Bear *et al.*, 1992), (Dowding *et al.*, 1993), le travail de Heeman présente une avancée significative. En effet, cette avancée est cependant à relativiser étant donné que les deux approches n'ont pas été testées sur le même corpus de test, et n'ont pas la même définition des différents phénomènes (en particulier le faux-départ et l'autocorrection). Par ailleurs, Heeman ne donne pas le pourcentage des extragrammaticalités imbriquées traités dans le cadre de son approche.

Une approche similaire à été proposé par (Stolke et Shriberg, 1996) dans l'objectif d'améliorer les résultats de la reconnaissance de la parole. Le résultat n'affiche pas une différence notable de performance seulement 0,02% quant à la perplexité du modèle de langage, le modèle augmenté affiche une perplexité supérieure de 1,8%.

2.1.1.32 Limites de l'approche de Heeman

1. **Insuffisance de l'information fournie par les POS tags :** l'utilisation des tags comme l'unique source de connaissance morphologique pour le traitement de certains phénomènes est trop limitative. En effet, dans certains cas nous avons besoin d'informations morphologiques détaillées afin de pouvoir analyser correctement un cas d'extragrammaticalité : personne, fonction syntaxique (sujet, objet pour les pronoms), etc. Prenons comme exemple la construction suivante : *prep + pronpers*. Cette construction est impossible si *pronpers* est sujet (to I) et elle est parfaitement grammaticale si l'élément de catégorie *pronpers* est objet (to it).
2. **Limitation syntaxique de N-grams :** cette limitation cache la dimension syntaxique et sémantique des extragrammaticalités. En effet, l'utilisation des N-grammes limite en la prise en considération du contexte morphologique à quelques mots alors qu'on a parfois besoin de contexte plus important pour pouvoir détecter une extragrammaticalité. Prenons les exemples suivants pour mettre au clair cette idée :

   It will take it. (41)

   It will take **it** is midnight. (42)

   We would have to do it. (43)

   We would have to **do** you think it is possible to do it. (44)

   Comme nous pouvons le remarquer dans l'énoncé 47, le mot *it* est considéré comme un objet et l'énoncé est justement considéré comme étant bien formé puisque le verbe *take* est un verbe transitif. Par contre, dans l'énoncé 48, le mot *it* peut être, à la fois, sujet et objet, d'une part, à cause de l'ambiguïté morphologique de cet item et d'autre part, à cause de sa situation entre deux syntagmes verbaux. Ainsi, nous avons besoin d'un dispositif qui prend en considération le contexte droit afin de désambiguïser cette structure syntaxique et décider que le premier syntagme est mal formé et qu'il s'agit, par conséquent, d'un faux départ. Les mêmes remarques s'appliquent à l'énoncé 50 où le verbe do peut appartenir à deux syntagmes.

3.2.4 L'approche à base de méta-règles syntaxiques de Mark Core

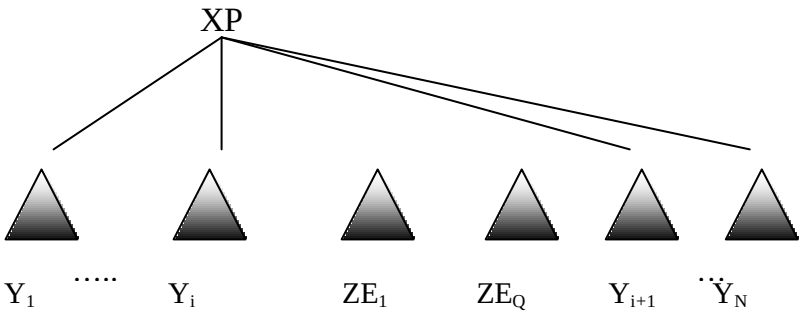
Ce travail est mené dans le cadre d'une approche générale de l'analyse robuste des dialogues au sein du groupe de dialogue de l'université de Rochester (Core, 1999). La particularité principale de ce travail est l'introduction des informations linguistiques (en particulier la syntaxe) dans le traitement

des extragrammaticalités d'une manière originale (différente de celle de SRI). En effet, Selon cette approche le traitement se fait en deux étapes :

1. **Détection des extragrammaticalités :** la détection des extragrammaticalité se fait avec un modèle de langage statistique (celui de Heeman, présenté dans la section précédente). La fonction principale de ce module est de détecter les extragrammaticalités et de proposer une **première** délimitation de chacune de ces extragrammaticalités.

2. **Analyse syntaxique :** la fonction du module d'analyse syntaxique est de donner une interprétation qui couvre la totalité des mots de l'énoncé d'entrée. Pour cela, il traite les extragrammaticalités détectées par le module statistique à l'aide de méta-règles dédiées spécialement à cette tâche. La différence principale entre le traitement dans cette phase et celui du Heeman est que le système considère non les relations entre les mots (comme c'est le cas dans l'approche de Heeman) mais plutôt les relations entre les structures syntaxiques qui dominent les mots. Deux types de méta-règles sont utilisés pour le traitement des extragrammaticalités :

i- **La méta-règle de la zone d'édition :** basée sur une liste de mots qui peuvent potentiellement constituer une zone d'édition ou une partie d'elle, la méta-règle de la zone d'édition détecte tous les segments susceptibles d'être une zone d'édition et déclenche directement la méta-règle de la zone d'édition.



i≥1, ≥ 2, Q >0

Figure 19. La règle de la zone d'édition proposée par (Core et Schubert, 1998)

Dans la figure précédente, *XP* peut correspondre à n'importe quel constituant d'un énoncé dont les sous-constituants peuvent être interrompus par une zone d'édition.

La méta-règle a été implantée au sein d'un algorithme de type *chart* en autorisant tous les syntagmes amorcés avant la zone d'édition potentielle d'apparaître après cette zone. En d'autres termes, la méta-règle permet d'analyser l'énoncé d'entrée sans considérer la zone d'édition.

ii- **La méta-règle des autocorrections et faux-départs :** la fonction principale de cette méta-règle est de délimiter une extragrammaticalité amorcée (par le module précédent) précisant le début et la fin des zones remplacées et remplaçantes puis, elle permet à

l'algorithme d'ignorer la zone remplacée et de considérer uniquement la zone remplaçante. Le schéma général de cette règle est présenté dans la figure suivante :

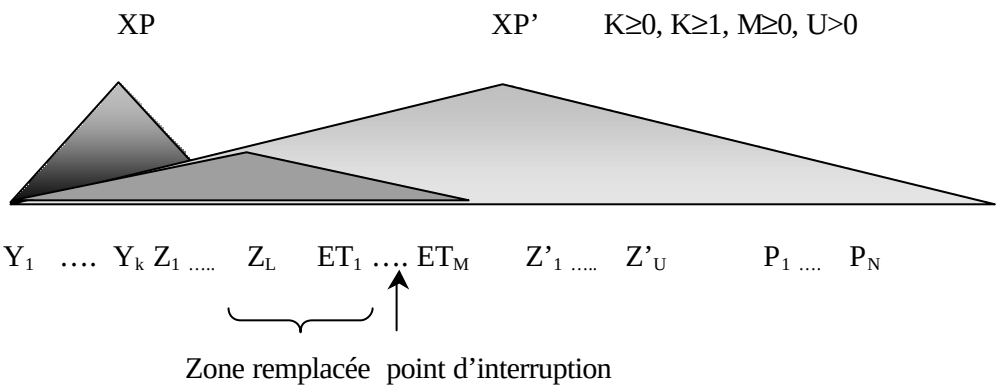


Figure 20. La méta-règle de traitement des autocorrections et faux départs

Dans la règle présentée ci-dessus, la nature des composantes *XP* et *XP'* n'est pas précisée, mais généralement, chaque composante est constituée d'un ensemble de syntagmes *Z* et *Z'* qui dépendent directement d'elle. Dans le cas d'autocorrection et faux-départs, les syntagmes Z_i et Z'_i tendent à être du même type. Ces méta-règles sont implantées selon le même principe que les méta-règles de la zone d'édition : les arcs qui se terminent avant la zone remplacée sont liés directement au début de la zone remplaçante, permettant ainsi de traiter l'énoncé en ignorant la zone remplacée ainsi que la zone d'édition qui peut la suivre.

L'utilisation généralisée des méta-règles pour tous les phénomènes nous semble difficile à justifier. En effet, le traitement d'une bonne partie de ces phénomènes (en particulier les extragrammaticalités lexicales, les répétitions et les autocorrections avec répétition) ne nécessite pas la mobilisation d'informations syntaxiques et peut être réalisé avec de simples techniques de reconnaissance de patrons qui sont, par ailleurs, plus restrictifs que les règles : généralement on considère l'identité des mots plutôt que leur catégorie morphologique, ce qui réduit considérablement les surgénérations.

Par ailleurs, l'intégration de méta-règles au sein d'un algorithme gauche droite de type *chart* ou autre ne permet pas la prise en considération des extragrammaticalités imbriquées étant donné que celles-ci

ont besoin de plusieurs passages[19]. En terme de calcul, l'ajout des méta-règles s'est révélé très coûteux (Core, 1999). En effet, le temps de traitement d'un énoncé avec un analyseur simple est de 0.36 secondes alors qu'avec un analyseur augmenté par les méta-règles, le temps est de 0.91. Autrement dit, l'ajout des méta-règles a augmenté le temps de calcul de trois fois approximativement.

Sur le plan des résultats, deux expériences ont été menées. Dans la première une ancienne version du système de Heeman a été utilisée. Les expériences de (Core, 1999) ont montré un avancement de 1.6% en terme de rappel mais une perte de 12,4% en précision. Réalisée avec une version plus récente du système de Heeman, la deuxième expérience a confirmé la légère amélioration du rappel avec l'ajout des méta-règles : le rappel a augmenté de 1,02%. Par ailleurs, cette expérience a montré une différence encore plus importante en terme de précision : la version augmentée affiche une précision inférieure de 40.33% au système de Heeman. Outre les critiques formulées précédemment, il nous semble que cette perte en précision est causée par la faible interactivité entre le module statistique de détection et l'analyseur symbolique. Par ailleurs, l'auteur avance la faible couverture de l'analyseur utilisé comme étant la première raison d'échec de traitement.

Des approches similaires à celle de Core ont été proposées par différents chercheurs. Par exemple, (McKelvie, 1998) propose une approche à base de méta-règles. Outre que les unités syntaxiques classiques, ses méta-règles considèrent deux catégories :

- **Les syntagmes d'éditions** (ED) qui sont les hésitations, bruits, exclamations, etc.
- **Les marqueurs discursifs** (AFF) qui correspondent à des mots comme *oui*, *ok*, etc. et qui marquent généralement le début et la fin d'un énoncé.

Les méta-règles utilisées sont assez simples généralement. Prenons, par exemple, la règle suivante :

$X \rightarrow X, ED, AFF$

Cette règle permet d'ignorer tous les éditeurs qui apparaissent après un constituant *X*.

Cette approche a été réalisée sur le Glasgow Maptask corpus, mais l'auteur ne donne pas de résultats expérimentaux.

[19] Ce point fera l'objet d'une discussion détaillée dans le premier chapitre de la quatrième partie de cette thèse.

4 Conclusion de la première partie

Dans cette partie, nous avons fait une revue générale des différentes propriétés linguistiques du langage oral utilisé en dialogue, ainsi que des principaux formalismes syntaxiques et sémantiques utilisés pour la représentation de ces différentes propriétés.

4.1 Bilan des Spécificités linguistiques du langage oral

Les principales spécificités de l'oral que nous avons passées en revue dans cette partie peuvent être résumés dans les deux points suivants :

- **Syntaxe :** nous avons vu les principales spécificités syntaxiques de l'oral par rapport à l'écrit.
- **Extragrammaticalités :** nous avons vu que les conditions de production de la parole en ligne impliquent des phénomènes d'hésitation, d'autocorrections, faux-départs, etc. qui sont propres à l'oral et qui nécessitent un dispositif particulier pour les traiter dans le contexte d'un système d'analyse linguistique du langage oral.

4.2 Bilan des formalismes utilisés pour la représentation de l'oral

Nous avons présenté dans cette partie deux formalismes grammaticaux que nous avons jugés représentatifs des travaux dans la littérature. Il s'agit du formalisme LTAG et ses dérivés ainsi que de la grammaire sémantique classique. Nous avons vu que ces deux approches ont des avantages et inconvénients opposés pour le traitement de l'oral. En effet, le formalisme LTAG est bien adapté pour le traitement profond et il permet la prise en considération des phénomènes syntaxique dans l'analyse. Par contre, ce formalisme ne permet pas une interaction suffisante avec la tâche du dialogue et a certaines difficultés à traiter des phénomènes comme les ellipses. A l'opposé, la grammaire sémantique permet facilement la prise en considération des phénomènes sémantiques mais elle échoue à prendre en considération de manière efficace de phénomènes syntaxique comme la négation. Par ailleurs, cette grammaire n'est pas définie formellement et n'a pratiquement pas de statut linguistique.

Dans cette partie, nous avons présenté quelques techniques d'analyse robuste et leur application au traitement du langage oral.

Afin de situer cette revue de la littérature dans le contexte de la problématique générale de notre thèse, nous avons jugé bon de dresser un bilan général qui synthétise les principaux points problématiques qui peuvent influencer nos choix futurs :

4.3 Bilan des approches d'analyse robuste du language oral

4.3.1 Les approches pour l'analyse syntaxique robuste

- **Les approches sélectives :** les approches sélectives semblent une bonne solution pour les problèmes de sous-génération de la grammaire du système (un problème qui résulte à la fois du presque inévitable manque de données ou des problèmes liés au bruit dans l'entrée). Le coût de ces avantages est généralement l'augmentation de la complexité algorithmique.
- **Analyse partielle :** l'approche d'analyse partielle semble bien adaptée aux besoins d'un système d'analyse linguistique du langage oral. En effet, cela donne une bonne au système par rapport aux différentes sources de manque de robustesse comme les extragrammaticalités, les erreurs de reconnaissance, etc.

4.3.2 Les approches pour le traitement des extragrammaticalités de l'oral

Plusieurs sources de connaissance ont été utilisées dans la littérature pour le traitement des extragrammaticalités. En voici les principales :

1. **Les informations structurales :** elles concernent l'identité de chaque mot et celles des mots qui le succèdent et suivent[20]. L'avantage de cette information est sa fiabilité et sa simplicité d'utilisation mais son utilisation est généralement limitée à la détection des répétitions. Certaines approches ont négligé cette source d'informations (Cori, 1997), (McKelvie, 1998), (Core, 1999), ce qui nous semble difficilement justifiable d'un point de vue pratique.
2. **Les informations morpho-syntaxiques :** elles concernent essentiellement les catégories morpho-syntaxiques des mots ou des segments (chunks) et leurs successions possibles. Par exemple, la succession de deux déterminants est jugée extragrammaticale et par conséquent, le cas est traité comme une autocorrection. Certains systèmes ont utilisé des règles plus complexes afin de modéliser des cas impliquant des constituants syntagmatiques. Dans ce genre de cas, des analyseurs classiques ont été faits pour assumer cette tâche. Ces règles ont généralement été implantées comme des méta-règles syntaxiques dans un module de post-traitement. Plusieurs remarques peuvent être formulées à propos de cette utilisation :
   i Dépendance du module de traitement des extragrammaticalités de l'analyseur syntaxique utilisé dans l'application, ce qui réduit considérablement sa portabilité.
   ii Coût élevé de traitement, puisque cela nécessite l'utilisation d'un analyseur syntaxique classique.

[20] C'est à dire le système vérifie si deux mots sont identiques ou pas sans se soucier de leurs catégories morphologiques respectives.

iii L'utilisation d'un analyseur syntaxique classique (pas robuste par rapport aux extragrammaticalité et aux erreurs de reconnaissance) peut être une source de certaines erreurs.

3. **Les informations acoustico-prosodiques :** il s'agit d'un ensemble d'informations de natures diverses comme la pause silencieuse et le contour mélodique qui ont été utilisées afin de segmenter l'entrée en constituants syntaxiques et par conséquent localiser le centre de l'extragrammaticalité dans l'énoncé.

4. **Les extragrammaticalités lexicales :** ces extragrammaticalités constituent une source importante pour la détection des extragrammaticalités supralexicales. Le problème est que certaines formes de ces extragrammaticalités (notamment les mots incomplets) ne sont pas reproduites par le système de reconnaissance. Cela rend l'utilisation de ces formes dans le traitement (comme l'a fait (Heeman, 1997)) une démarche irréaliste.

Partie II : Etude des phénomènes grammaticaux et extragrammaticaux du langage oral

0 Introduction de la deuxième partie

Après avoir fait une revue générale des propriétés du langage oral, des différents formalismes qui peuvent être utilisés pour sa représentation ainsi que des différentes approches dans le domaine de l'analyse syntaxique robuste du langage parlé, nous allons dans cette partie présenter notre contribution à l'étude de l'oral sur deux axes :

- Analyse du *Trains Corpus* dont nous avons extrait environ 6000 cas d'extragrammaticalités lexicales et 928 cas d'extragrammaticalités supralexicales.
- Modélisation grammaticale de l'oral. Sur ce plan nous avons contribué à deux niveaux :
  i. Formalisation de la grammaire sémantique et sa représentation en tant qu'une grammaire d'arbre au sein de laquelle différents niveaux d'unités peuvent être respectés.
  ii. Proposition du formalisme Sm-TAG qui intègre, à côté des informations sémantiques, des informations syntaxiques explicites.

1 Chapitre II.1 : Analyse des extragrammaticalités du langage oral dans le *Trains corpus*

1.1 Introduction

Dans ce chapitre, nous nous proposons pour faire une étude théorique basée sur la considération des différentes sources linguistiques susceptibles de jouer un rôle dans la représentation et le traitement des extragrammaticalités avec une attention particulière sur la dimension syntaxique de ces phénomènes. L'aspect prosodique, bien qu'important, ne sera pas abordé dans notre étude. Ceci est dû à plusieurs raisons :

- Théoriques : essentiellement dû au fait que cet aspect nous semble bien étudié par les autres chercheurs (Nakatani et Hirshberg, 1994), (Lickley, 1994), (Shriberg, 1994).
- Pratiques : Outre le fait que l'analyse linguistique robuste est l'objectif principal de notre thèse. La limitation de notre étude aux aspects linguistiques nous permet d'aller plus loin dans l'analyse du rôle de ceux-ci qui sont moins explorés que celui de la prosodie.

1.2 Le corpus d'étude

1.2.1 Sélection du corpus

Au début de notre étude, nous avons essayé de trouver un corpus qui contient un nombre raisonnable d'extragrammaticalité et dont l'annotation est faite de manière suffisamment fine pour nous permettre d'observer les différentes propriétés de ces phénomènes. Malheureusement, nous n'avons pas réussi à trouver un tel corpus pour le français. Ainsi, nous avons décidé de travailler sur l'anglais en raison de la disponibilité de sources linguistiques importantes pour cette langue.

Après avoir effectué différentes recherches dans notre entourage aussi bien que sur Internet nous avons réussi à collecter des extraits de trois corpus considérés comme étant des corpus standards dans le domaine du dialogue oral spontané orienté par la tâche. Il s'agit du *Trains Corpus* (Heeman et Allen, 1995), du Corpus ATIS (Hemphill, 1990) et du *Switchboard Corpus* (Godfrey *et al.*, 1992). Après avoir comparé les trois corpus, nous avons opté pour le *Trains Corpus* pour les raisons suivantes :

- La finesse d'annotation : l'annotation du *Trains Corpus* est la plus fine des trois corpus notamment en ce qui concerne les événements liés aux extragrammaticalités (clicks, silences, etc.).

- Bien que les corpus ATIS et Switchboard aient fait l'objet d'études portant sur les extragrammaticalités, le Trains corpus est celui qui nous permet d'effectuer les meilleures comparaisons de notre travail avec les travaux précédents (notamment en terme de qualité de traitement). En effet le Swichboard corpus et l'ATIS corpus ont fait l'objet d'études essentiellement descriptives (Hirschberg et Nakatani, 1994), (Meteer *et al.*, 1995) alors que le *Trains Corpus* a fait l'objet de deux études clés pour notre travail : celle de (Heeman, 1998), et celle de (Core, 1999).
- Disponibilité : la totalité de ce corpus est disponible gratuitement sur Internet[21] aussi bien qu'à travers la *Linguistic Data Consortium* (LDC) (contrairement au corpus ATIS qui n'était disponible que partiellement).

1.2.2 Validité de nos observations dans le Trains Corpus

L'une des premières questions que nous nous sommes posées au début de notre travail sur le *Trains Corpus* était la validité de nos observations sur d'autres corpus, en particulier en ce qui concerne les phénomènes complexes. Ainsi, nous avons essayé de vérifier les occurrences des phénomènes complexes (comme les extragrammaticalités imbriquées que nous allons voir plus loin) dans un autre corpus. Pour ce faire, nous avons procédé à une annotation informelle des extragrammaticalités dans une dizaine de dialogues extraits du *Swicboard corpus*. Les résultats de notre annotation nous ont permis d'observer une similarité des phénomènes dans les deux corpus tant simples que complexes (y compris les occurrences multiples des extragrammaticalités[22]). Par ailleurs, nous avons observé informellement (que ça soit dans nos interactions personnelles ou dans dialogues oraux diffusés à travers les médias audio ou audiovisuels) les différentes formes d'extragrammaticalités que nous avons trouvées dans *le Trains Corpus*.

1.2.3 Présentation du Trains Spoken Dialog Corpus

Le *Trains Spoken Dialog Corpus* (désormais le *Trains corpus*) est le corpus que nous avons utilisé dans notre analyse théorique des extragrammaticalités. Il s'agit d'un corpus qui a été collecté par Peter Heeman et James Allen (Heeman et Allen, 1995) à l'université de Rochester aux Etats Unis. La tâche de ce corpus est la négociation de transport de marchandises via le chemin de fer. Notre choix de ce corpus a été motivé par la fréquence relativement élevée des extragrammaticalité ainsi que la complexité des phénomènes observés (notamment à cause de la complexité de la tâche de dialogue) d'une part et d'autre part à cause de la bonne qualité du corpus tant d'un point de vue collecte que transcription. Les propriétés clés de ce corpus sont présentées en détail dans les points suivants :

---

[21] Ce corpus est disponible en ligne à l'URL suivant :

http://www.cs.rochester.edu/research/cisd/resources/trains.html

[22] Pour des exemples de ces cas complexes dans le *Swichboard Corpus* le lecteure peut consulter (Meteer *et al.*, 1995) page 15.

1. **La technique de collecte :** ce corpus a été collecté selon la technique du magicien d'Oz. Ainsi, deux personnes sont impliquées à chaque collecte de données. La première joue le rôle de la machine la deuxième joue le rôle du client. Un coordinateur était aussi présent à chaque enregistrement afin de surveiller la qualité du travail.
2. **Haute qualité d'enregistrement du signal de la parole :** cela permet d'utiliser ce corpus pour l'entraînement des modèles acoustiques des systèmes de reconnaissance de la parole, et surtout cela aide à augmenter la qualité de la transcription notamment concernant la transcription des hésitations, des mots incomplets, des silences, etc. qui sont des indices précieux dans l'étude des extragrammaticalités.
3. **Sujets :** les sujets qui ont joué le rôle du système sont des experts en informatique familiers avec la tâche du dialogue. Par contre, les sujets qui ont joué le rôle du client sont généralement des *naïfs* non familiers aux systèmes de dialogues homme-machine. 34 sujets ont participé à la collecte du corpus et ont formé 25 paires d'interlocuteurs.
4. **La tâche des dialogues :** la tâche des dialogues est la négociation du transfert de marchandise d'une ville à l'autre. Cette tâche a été décrite sous formes de scénarios dont le nombre est de 20. Le plan général correspondant aux tâches à accomplir est présenté dans la figure 37.
5. **Transcription :** tous les mots ont été reproduits dans la transcription avec le respect de leur orthographe : les mots normaux, les mots incomplets, les mots amalgamés, etc. Outre les mots, certains indicateurs phonétiques ou prosodiques ont été reproduits dans la transcription comme : les silences, les clicks, les bruits, les rires, etc.
6. **Taille :** la partie distribuée publiquement du *Trains Corpus* comporte 93 dialogues, 52000 mots (approximativement) et environ 5300 tours de parole.

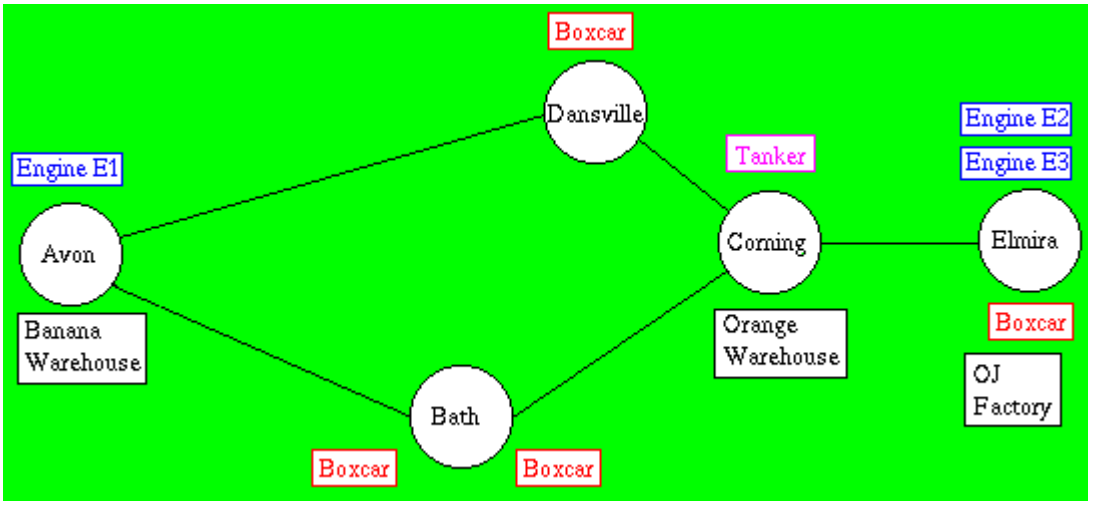


Figure 21. Le plan utilisé pour la collection du *Train corpus*

1.3 Annotation des données

1.3.1 Proposition d'un schème d'annotation des extragrammaticalités

Nous avons adopté un schème d'annotation qui consiste à extraire les informations nécessaires et suffisantes à la modélisation des phénomènes que nous cherchons à étudier. La particularité de notre démarche consiste à établir un système d'annotation différent pour chaque ensemble de phénomènes qui partagent les même propriétés. Ainsi, nous avons adopté trois systèmes différents qui correspondent aux principales formes d'extragrammaticalités observées. Notre schème d'annotation n'a pas été proposée *a priori*. En effet, tout d'abord, nous avons analysé une partie du corpus afin d'observer les tendances générales. Sur la base de cette observation, nous avons ensuite construit une première version de la méthode d'annotation. La version définitive a été faite en enrichissant la première version au fur et à mesure de l'analyse du corpus. Les détails des schèmes seront donnés dans les paragraphes suivants avec la présentation du processus d'annotation.

L'annotation des données consiste à associer les extragrammaticalités observées dans le corpus aux étiquettes correspondantes que nous avons adoptées dans notre schème d'annotation. Nous avons suivi trois procédures différentes une pour chacun des groupes de phénomènes pour lequel nous avons proposé une méthode d'annotation spécifique.

1.3.2 Les extragrammaticalités lexicales

L'objectif de l'annotation des extragrammaticalités lexicales est de repérer toutes les formes de ces phénomènes. L'annotation dans cette phase porte uniquement sur les extragrammaticalités lexicales en occurrences isolées. C'est-à-dire indépendamment d'une extragrammaticalité supralexicale, les extragrammaticalités lexicales qui apparaissent au sein d'une extragrammaticalité supralexicale étant considérés comme un élément de cette dernière.

Les fragments de mots ont été négligés à la fois lorsqu'il s'agissait d'une occurrence simple d'incomplétude de mot ou d'une occurrence au sein d'une autre extragrammaticalité. D'une part, parce que le traitement de ce phénomène est trivial (dans une application ciblée, il suffit de filtrer les mots inconnus) et d'autre part, dans le cas d'une entrée orale (qui est l'application que nous visons derrière notre étude des corpus transcrits), les mots incomplets ne sont pas reproduits par le système de reconnaissance de la parole.

2.1.1.33 Annotation des hésitations

L'annotation des hésitations est assez simple, il s'agit de faire la liste de toutes les formes d'hésitations observées dans le corpus. Les résultats de notre analyse sont présentés dans le tableau suivant :

Hésitation	Nb Occurrences
um	1013
uh	1171
mm	337
mm-hm	301
hm	293
oh	282
huh	49
uh-huh	44
ooh	11
ah	11
Total	**3512**

Tableau 2. Les hésitations observées dans notre corpus et leurs fréquences

Après avoir fait quelques opérations simples de calcul, nous avons trouvé que les hésitations constituent 6,75% des mots dans notre corpus et que 66,26% des énoncés contiennent des hésitations.

2.1.1.34 Annotation des amalgames

L'amalgame est un phénomène grammatical dont certaines formes sont propres à l'oral. Nous avons jugé bon d'inclure l'analyse de ces phénomènes dans notre étude étant donné que dans certains cas les occurrences des amalgames ont un effet direct sur le traitement des extragrammaticalités comme la succession d'une expression amalgamée et de la même expression en forme standard : *I'll I will* (dans ce cas, on peut dire que l'objet de l'extragrammaticalité est la correction de l'amalgame qui dénote un niveau de conversation informel).

Dans cette étape, notre travail a consisté à faire la liste de toutes les formes d'amalgames observées dans le corpus et les associer à leurs formes standards qui sont utilisées tant à l'oral qu'à l'écrit. La liste complète des mots trouvés avec leurs fréquences est présentée dans le tableau 3 :

Amalgames et mots oraux	Forme standard	Construction	Occurrences
Aren't	Are not	Verbe adverbe	7
Avon's	Avon is	Prop verbe	1
Can't	Can not	Verbe adverbe	56
Could've	Could have	Vaux verbe	4
Didn't	Did not	Verbe adverbe	13
Doesn't	Does not	Verbe adverbe	37
Don't	Do not	Verbe adverbe	91
Hadn't	Had not	Verbe adverbe	2
Hasn't	Has not	Verbe adverbe	1
Here's	Here is	Adverbe verbe	69
Wasn't	Was not	Verbe adverbe	5
I'd	I would	Pronpers vaux	61
I'll	I will	Pronpers vaux	143
I'm	I am	Pronpers verbe	123
it'd	It would	Pronpers vaux	6
It'll	It will	Pronpers vaux	83
It's	It is	Pronpers verbe	194
I've	I have	Pronpers verbe	7
Gotta	got to	Verbe prep	14
Let's	Let us	Verbe Pronpers	156
Long's	Long is	Adverbe verbe	2
One's	One is	Pron verbe	5
That'd	That would	Pron vaux	4
That'll	That will	Pron vaux	100
That's	That is	Pron verbe	351
Them's	Them is	Pronpers verbe	1
There'd	There would	Adv vaux	1
There's	There is	Adv verbe	65
They'll	They will	Pronpers vaux	18
They're	They are	Pronpers verbe	15
Wanna	I want	Pronpers verbe	61
We'll	We will	Pronpers vaux	135
We're	We are	Pronpers verbe	130
We've	We have	Pronpers verbe	8
Who's	Who is	Pron verbe	2
Won't	Will not	Vaux adv	24
Wouldn't	Would not	Vaux adv	17
You'd	You would	Pronpers vaux	14
You'll	You will	Pronpers vaux	28
You're	You are	Pronpers verbe	69
You've	You have	Pronpers verbe	21
		Somme des cas	**2212**

Tableau 3. Formes d'amalgames et leurs fréquences

Selon nos calculs, 4,22% des mots de notre corpus correspondent à des amalgames et 41,47% des énoncés contiennent une occurrence d'un amalgame.

Comme le montre le tableau 3, les différentes formes d'amalgames observées dans notre corpus correspondent à des constructions verbales, en particulier les constructions pronom verbe qui occupent une place largement dominante parmi les occurrences des autres formes. La répartition des occurrences entre les différentes constructions est présentée dans la figure 38.

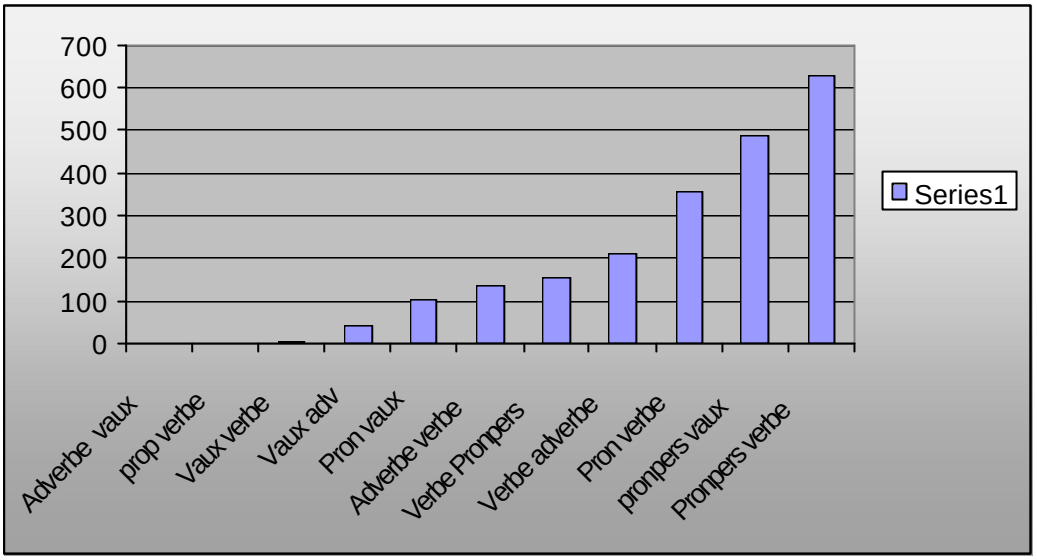


Figure 22. Les différentes constructions d'amalgames et leurs fréquences

Comme le montre la figure précédente, les trois constructions les plus dominantes sont les constructions pronom personnel verbe, pronom personnel verbe auxiliaire et pronom verbe. Nous remarquons aussi que les onze constructions observées dans notre corpus impliquent un verbe ou une forme verbale.

2.1.1.35 Annotation des mots oraux

Dans notre corpus, nous avons observé des mots qui sont des équivalences elliptiques de mots *standards* que nous avons appelé les mots oraux. Bien que ces mots ne soient pas une partie intégrante des extragrammaticalités (il s'agit en effet d'un phénomène grammatical propre à l'oral), nous avons jugé bon de les inclure dans notre étude étant donné que ces mots entrent dans la construction de certaines extragrammaticalités et ont un effet direct sur leur traitement (comme dans la répétition d'un mot oral et de sa forme standard : yeah yes). Les deux exemples les plus courants que nous avons observés sont :

- Les équivalents informels de *yes* (238 occurrences) : *yeah* (235 occurrences) et *yep* (108 occurrences).
- La version orale de *because* (68 occurrences) : *cause* (68 occurrences).

Ces statistiques nous montrent que les versions standards et les versions informelles (propres à l'oral) sont tous les deux utilisés malgré l'avantage relativement léger des mots informels.

1.3.3 Les extragrammaticalités supralexicales

2.1.1.36 Annotation des répétitions et autocorrections

Le schème d'annotation des répétitions et des autocorrections est essentiellement inspiré des travaux de (Bear, 1994) avec certaines modifications. En voici les principales étiquettes :

M_x	Mots identiques
R_x	Remplacement
E	Editeurs (silence, hésitation, mots incomplets)
X	Mots neutres

Tableau 4. Signes utilisés pour l'étiquetage des répétitions et des autocorrections

Ainsi, toutes les répétitions et autocorrections observées sont étiquetées sous forme de patrons réunissant leurs différents éléments. Contrairement aux approches précédentes, nous avons abordé l'étiquetage du corpus avec le minimum de préjugés. Ainsi, nous avons adopté deux méthodes d'étiquetage :

- La première, **globale**, consiste à étiqueter les phénomènes tels qu'ils apparaissent dans le corpus, ce qui nous a permis de considérer des patrons représentant des successions d'extragrammaticalités (des occurrences de plusieurs phénomènes en un même énoncé). Voici un exemple d'énoncé étiqueté (selon le format global) :

  (**…**) <sil> do I <sil> I need two <sil> do I need two <sil> engines for the (**...**)

  M1 M2 E M22 M3 M4 E M12 M23 M32 M42

  Nous remarquons que dans l'énoncé précédent les extragrammaticalités ont été annotés tels qu'elles sont sans aucune segmentation *a priori* ce qui nous permet d'observer la relation entre les deux extragrammaticalités qui se trouvent dans cet énoncé.

- La seconde, **locale,** similaire à celle de (Bear, 1994), et de (Heeman, 1994), consiste à considérer chacune des occurrences à part. Le résultat de notre étiquetage consiste en 48 patrons (locaux) dont les plus fréquents sont représentés dans les tableaux 5 et 6. Il faut distinguer entre la forme de surface (les patrons) et la forme linguistique qui indique la nature linguistique des éléments remplacés. Ces représentations superficielles ont été enrichies par des annotations des modifications syntaxiques dans les autocorrections. Par exemple, l'énoncé : okay so that'll take <sil> so that'll be seven a.m. (d93-10.5, utt12) est annoté de la manière suivante :

1. Un patron correspondant à la structure superficielle de l'autocorrection est construit : M1M2M3R1 M1M2M3R1'[23].
2. Une paire de transition correspondant aux catégories des faux-départs qui sont impliqués dans le phénomène.

1.3.3.1.1 Les répétitions

Par répétition, nous entendons la reprise d'un mot ou d'un ensemble de mots pas celle d'un segment différent avec le même sens (la paraphrase). Ainsi, des cas comme l'énoncé 56 ne sont pas considérés comme des répétitions. Puisque les deux segments repris ne sont pas parfaitement identiques en terme de mots.

(…) **engine E one wasn't** <sil> maybe **it wasn't** the best thing. (d93-19.5, utt52) (45)

Cette règle n'est cependant pas absolue. En effet, elle ne permet pas toujours de décider la nature du phénomène en cas d'ambiguïtés formelle entre une autocorrection et une répétition. Dans notre corpus, deux types d'ambiguïtés ont été observés :

1. Répétition avec l'insertion d'un mot : il s'agit de l'insertion d'un mot avant la zone répétée. Dans ce cas, nous sommes devant une ambiguïté puisque ces cas peuvent être considérés comme une répétition ou comme une autocorrection par insertion. Selon la nature des mots insérés, nous avons distingué deux cas :

i. Insertion d'un mot qui peut être un éditeur comme dans l'exemple suivant :

(…) let's see maybe **it would** yeah **it would** (…)
(d93-26.2, utt41) (46)

Dans ce cas, nous pouvons considérer qu'il s'agit d'une autocorrection par insertion de *yeah* qui est utilisé pour renforcer le sens du segment et en même temps, nous pouvons considérer qu'il s'agit d'une répétition puisque nous avons deux segments identiques qui sont séparés par un éditeur. Nous avons décidé de classer ces cas avec les répétitions, d'une part à cause de la forte ressemblance avec les répétitions normales avec une zone d'édition et d'autre part, à cause du rôle secondaire de la modification sémantique apportée par l'insertion de ce genre de mots.

ii. Insertion d'un mot normal : il s'agit généralement de l'insertion d'un modifieur (adverbe, adjectif, etc.) avant la zone répétée. A titre d'exemple, examinons le cas suivant :

The probably **the** trip from Avon to Corning takes (…) (d93-19.4, utt29) (47)

[23] Nous présupposons que l'amalgame I'll a déjà été résolue.

Dans le segment ***The*** *probably* ***the,*** si nous considérons le mot *probably* comme étant un mot neutre, nous pouvons juger le cas comme une répétition avec un mot neutre entre les deux segments répétés et le cas peut ainsi être annoté avec le patron M1XM1. Par contre, si nous considérons la modification sémantique apportée par l'adverbe *probably*, il nous semble clair qu'il s'agit plutôt d'une autocorrection.

2. **Répétition avec la suppression d'un mot :** prenons l'exemple suivant :

(…) so **we *just* need to get** um <sil> let's see <sil> **we need to get** um <sil> to <sil> Dansville <sil> two boxcars of oranges (utt34, d93-11.2) (48)

L'extragrammaticalité dans l'énoncé précédent, peut être considéré une répétition avec un mot inconnu dans l'une des deux parties du patron comme : M- X M-. Ce qui renforce le choix de la répétition est que le mot supprimé joue un rôle sémantique mineur et les segments remplacé et remplaçant ont pratiquement le même sens. De même, ce phénomène peut être considéré comme une autocorrection avec suppression étant donné que la partie remplacée et la partie remplaçante de l'extragrammaticalité ne sont pas parfaitement identiques.

Dans notre corpus nous avons observé 256 cas de répétitions répartis sur 12 patrons différents. Les différents patrons observés ainsi que leurs fréquences sont présentés dans le tableau suivant :

Patron	%
M1 ed M1	43,95
M2 ed M2	25,82
M1 M1	10,98
M4 ed M4	4,39
M3 ed M3	3,46
M5 ed M5	3,29
M1 X M1	3,29
M2 X M2	1,64
M1 ed M3 ed M3	1,09
M6 M6	0,05
M4 M4	0,05
M1 ed M5 ed M5	0,05

Tableau 5. Les patrons de répétition avec leurs pourcentages

Comme nous pouvons le remarquer dans ce tableau, il existe une tendance générale selon laquelle la fréquence d'un patron est inversement proportionnelle à sa taille. Autrement dit, plus le patron est petit plus il est fréquent et vice versa.

1.3.3.1.2 Les autocorrections

Nous avons observé trois procédés d'autocorrection dans notre corpus :

- **L'insertion d'un mot :** comme nous avons vu dans la section précédente, il s'agit d'une modification sémantique apportée à un segment par l'insertion d'un mot au début ou au à l'intérieur de ce segment.
- **Le remplacement d'un mot :** dans ce cas, on remplace un mot par un autre souvent de la même catégorie ou dont le rôle fonctionnel est assez proche (comme : cardinal ou déterminant). Le remplacement est parfois accompagné par la reprise d'une série de mots comme dans l'énoncé suivant :

  yeah I need to ship <sil> one boxcar of **bananas** <sil> one boxcar of **oranges** <sil> and one tanker of OJ <sil> to Bath (utt2, d93-11.3) (49)

- **La modification de l'ordre des mots :** ce procédé est particulièrement utilisé pour remplacer une construction verbale affirmative par une construction interrogative comme dans l'énoncé suivant :

  I don't know if **that's is that** the maximum number <sil> possible <sil>

  (utt27, d93-8.3) (50)

Notre corpus contient 241 cas d'autocorrections répartis sur 35 patrons. Voici les 15 patrons les plus fréquemment observés avec leurs fréquences :

Patron	%
R1R1	24,71
M1R1	8,64
M2R1	6,74
R1edR1	6,74
R1M1	6,74
M2R1Xed	5,61
M1R1M2	3,37
R1R2	3,37
M1M2R1M3M4R2M5 M1M2R1'M3M4R2'M5	2,24
M1R1R2	2,24
M2R1M3	2,24
M3R1	2,24
R1M4 R1M4	1,12
R1R2M1	1,12
R1XR1	1,12
M4XM4	1,12

Tableau 6. Les quinze patrons d'autocorrection les plus courants avec leurs fréquences

La tendance observée dans les répétitions est aussi confirmée avec les autocorrections : la fréquence d'un patron est inversement proportionnelle à sa taille.

2.1.1.37 Annotation des faux-départs

Comme nous avons vu dans la première partie de cette thèse**,** les faux-départs consistent à abandonner ce que le locuteur vient de dire et à recommencer à nouveau. Ce processus d'abandon segmente l'énoncé en plusieurs zones ayant des fonctions différentes.

Dans notre corpus, nous avons observé 272 cas de faux-départs dont 25% se trouvent dans des faux-départs multiples. Ainsi, nous avons utilisé des règles pour annoter ces phénomènes. Le schéma général de ces règles est présenté dans la figure suivante :

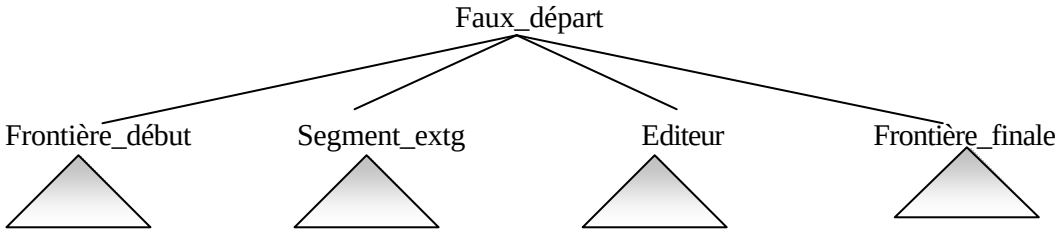


Figure 23. Schéma général des autocorrections

Comme nous pouvons le remarquer dans la figure précédente, les zones impliquées dans ces règles sont les suivantes : la frontière de début, le segment extragrammatical, la zone d'édition et la frontière finale. Avant de présenter les propriétés de chaque zone dans le schéma, nous allons commencer par la présentation des relations entre ces unités.

1.3.3.1.3 Analyse des relations de dépendance entre les zones clés du faux-départ

Les dépendances syntaxiques entre les différentes zones au sein d'un faux-départ ont un impact important sur la détection de ceux-ci. En effet, dans certains cas, l'élément situé après la rupture (qui peut être marquée par une zone d'édition ou par la prosodie seulement) peut être vu comme un complément naturel du dernier syntagme de l'extragrammaticalité. Examinons le schéma suivant pour mettre au clair cette idée :

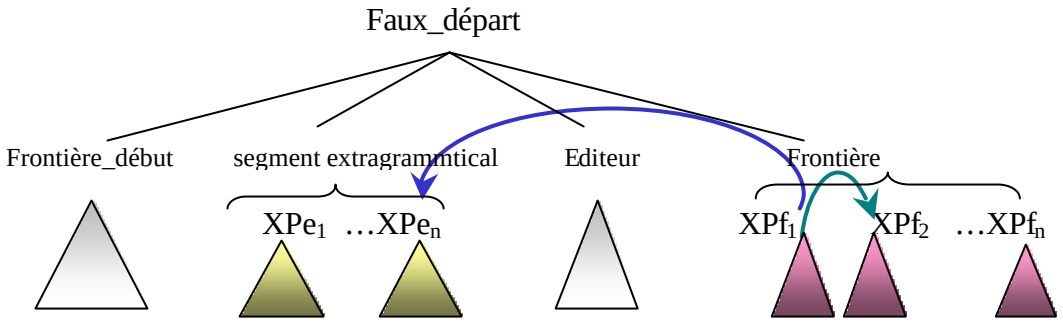


Figure 24. Schéma général des relations de dépendances entre les constituants du segment extragrammatical et de la frontière finale

Comme nous pouvons le voir dans la figure précédente, il peut y avoir une ambiguïté de l'attachement du premier syntagme de la frontière XPf_1 qui peut être vu comme dépendant/dominant du dernier syntagme du segment extragrammatical XPe_n aussi bien que du deuxième syntagme de la frontière XPf_2.

Prenons l'énoncé suivant à titre d'exemple : *So it is not gonna be going to <sil> the easiest way is to go to Bath or Corning*. Prenons l'arbre d'analyse simplifié de l'énoncé pour montrer l'ambiguïté de dépendance :

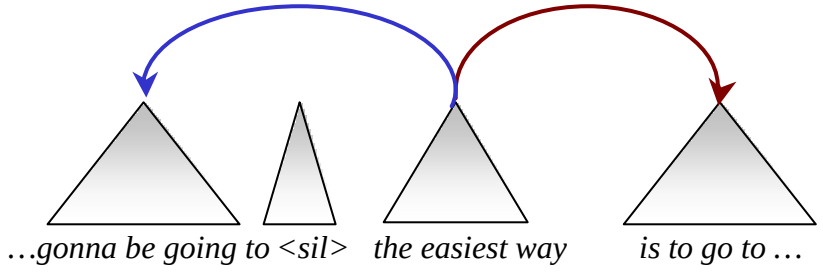


Figure 25. Arbre montrant l'ambiguïté de dépendance du premier syntagme d'une extragrammaticalité

Comme nous pouvons voir dans l'arbre d'analyse précédent le syntagme nominal *the easiest way* peut être attaché au dernier syntagme du segment extragrammatical. A ce moment là il est considéré comme un objet du prédicat verbal *going*. De même, ce syntagme peut être attaché au prédicat verbal *is* à droite et à ce moment là il joue le rôle de sujet. Ainsi, dans le contexte d'un système de traitement des extragrammaticalité, la résolution de l'ambiguïté d'attachement des segments frontières est un facteur décisif pour la détection d'un faux-départ.

1.3.3.1.4 Analyse des zones clés d'un faux-départ

1. Frontière de début

La délimitation de la frontière de début d'un faux-départ est nécessaire afin d'établir la limite gauche de cette extragrammaticalité. Plusieurs sources d'informations sont aussi utilisées afin de délimiter cette frontière :

- **La marque de début de l'énoncé :** les cas qui commencent au début de l'énoncé ont automatiquement leur frontière gauche qui est délimitée par ce début. Selon nos statistiques, cette source d'information est particulièrement utile puisque 71% des faux-départs sont situés au début de l'énoncé et 12% (c'est-à-dire 38% du reste) se trouvent dans des faux-départs multiples qui commencent au début de l'énoncé.

- **Des informations diverses :** différents types d'indices linguistiques, sémantiques et discursifs sont utilisés pour marquer le début d'un faux-départ. Ainsi, des mots comme les adverbes ou les hésitations sont considérés comme des indices particulièrement utiles dans la délimitation de la frontière gauche d'un faux-départ.

2. Segment extragrammatical[24]

Nous avons établi une typologie syntaxique générale des segments extragrammaticaux afin d'observer leur degré de régularité. Dans notre corpus, nous avons pu distinguer entre deux types de segments extragrammaticaux :

1. Des segments extragrammaticaux composés d'un seul syntagme auquel il manque un ou plusieurs mots.
2. Des segments composés de plusieurs syntagmes dont le dernier est incomplet.

Dans les deux cas précédents, le dernier syntagme du segment extragrammatical est la partie qui détermine sa nature. Ainsi, dans notre typologie nous nous sommes concentrés uniquement sur ce dernier syntagme. Le schème d'annotation que nous avons adopté consiste à annoter les syntagmes complets au sein du segment extragrammatical par le label : n(XP) où *n* (*n* ³ 0) est le nombre des syntagmes et *XP* est un syntagme de nature quelconque. Par exemple, le segment extragrammatical : pronpers Vpres infto (I want to) est présenté comme :

1(XP) vpres infto

Selon nos calculs, les valeurs de *n* vont entre 0 et 4 syntagmes. Le pourcentage des constructions où *n* est égal à zéro est de 15,78% et la moyenne de *n* est de 1,64 syntagmes. Nous avons observé au total 29 constructions dont 84,49% se terminent par un verbe ou un verbe suivi par un argument (préposition, adverbe, etc.).Voici le tableau général des principales constructions observées :

[24] Ces segments sont appelés : segments_extg dans la règle schématique pour des raisons de concision.

Structure	Pourcentage
n(XP) v	51,92
n(XP) v infto	19,23
n(XP) det	9,61
n(XP) pron	7,69
n(XP) v adv	5,769
n(XP) v pronpers	3,84
n(XP) vaux vinf	3,84
n(XP) v prep	3,84
n(XP) v det	3,84
n(XP) coord adv	3,84
n(XP) adv prep	3,84

Tableau 7. Les principales structures des segments extragrammaticaux dans les faux-départs et leurs fréquences

Par ailleurs, le problème de la détection des segments extragrammaticaux est que les critères de décision dans ce cas ne sont pas toujours absolus. Ainsi, nous pouvons distinguer entre deux types de segments extragrammaticaux :

i- **Des segments absolument extragrammaticaux :** il s'agit de segments qui sont jugés comme extragrammaticaux quel que soit le contexte dans lequel ils apparaissant. Généralement, ce sont les formes les plus simples des faux-départs, comme les occurrences isolées de déterminants, de prépositions, etc.

ii- **Des segments relativement extragrammaticaux :** il s'agit, dans ce cas, de segments qui sont parfaitement grammaticaux dans certains contextes et qui sont extragrammaticaux dans d'autres. Ce sont généralement des formes impliquant des structures syntaxiques complexes comme dans le segment :

pronpers + vpres + infto

Ce segment est considéré comme étant parfaitement grammatical s'il est suivi par un verbe infinitif mais il est jugé extragrammatical s'il est suivi d'un syntagme nominal par exemple.

3. Zone d'édition

La modélisation de la zone d'édition est similaire à ce que nous avons vu dans les patrons, la différence c'est que dans certains cas l'existence de la zone d'édition est obligatoire pour la considération d'un segment comme étant extragrammatical. Parfois, le type même de cette zone est décisif pour juger qu'il s'agit d'une extragrammaticalité. Soit l'exemple suivant :

but but um it was okay w- what um we only need one boxcar of OJ right (d93-18.4, utt76) (**51**)

Dans cet exemple, nous sommes devant deux possibilités d'analyse pour le faux départ souligné :

- Considérer que les deux verbes de l'énoncé *was* et *need* appartiennent à la même construction verbale et par conséquent considérer le pronom *what* comme un pronom objet.
- Considérer que l'énoncé contient deux prédicats verbaux : *was* et *need :* un premier (celui de *was*) avec une construction extragrammatical et le second (*celui de need*) correspond la zone reprise.

L'existence de la zone d'édition *okay w- what um* est le seul élément qui permet, dans ce cas, de faire le choix entre les deux interprétations et par conséquent de trancher en faveur de la deuxième.

4. Frontière finale

La frontière finale peut consister en un seul mot, un seul segment ou même une série de segments. Cette frontière a une double fonction, d'une part, elle permet de marquer l'étendue d'une extragrammaticalité et d'autre part, elle sert à réduire la surgénération d'une règle en contraignant le contexte droit d'un segment extragrammatical.

2.1.1.38 Annotation des incomplétudes

Le schéma d'incomplétude est assez similaire à celui du faux-départ. En effet, la seule différence entre les deux est que, dans le faux-départ, la plupart des cas commencent au début de l'énoncé (et ont donc leur frontière gauche qui est délimité *a priori*) alors que leur frontière droite est à délimiter. Pour les incomplétudes, il est rare de trouver un segment qui commence au début de l'énoncé mais, par définition, les incomplétudes n'ont pas un contexte droit et donc pas de zone d'édition, ce qui résout une partie du problème. Le schéma général des règles d'incomplétudes est présenté dans la figure suivante :

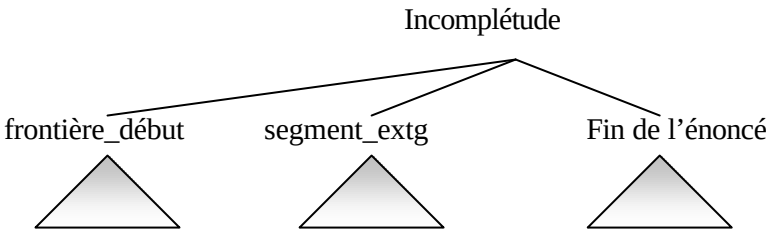


Figure 26. Schéma général des incomplétudes

Dans notre corpus, nous avons observé 83 cas d'incomplétude. Pour annoter ces cas, nous avons adopté le schéma général que nous avons utilisé avec les faux-départs : n(XP). Même si la procédure

d'annotation est similaire à celle des faux-départs, deux précisions liées à la nature des incomplétudes méritent être citées :

1. Nous avons compté tous les syntagmes observés y compris les syntagmes répétés ou corrigés. Par exemple, dans l'énoncé : *I think I think the bannana are already there because like* nous comptons les syntagmes *I* et *think* chacun deux fois.
2. Par ailleurs, nous nous sommes limité dans notre analyse aux mots complets étant donné que nous ne pouvons pas associer une catégorie grammaticale aux mots incomplets dont l'identité est généralement inconnue (les sujets prononcent les premières lettres d'un mot et en général cela ne suffit pas à l'identifier).

D'après nos calculs, la valeur moyenne de *n* (le nombre moyen des syntagmes qui précèdent le segment final) est de 3,70. Nous n'avons pas observé de cas ou le nombre des syntagmes qui ont précédé le syntagme final est 0. Par ailleurs, le nombre le plus large de syntagmes que nous avons observés était 10 (un cas unique). Comme nous pouvons le remarquer, ces valeurs sont supérieures à celles observées avec les faux départs (où les valeurs de *n* variaient entre 0 et 4 et la moyennes des syntagmes précédant le syntagme final était de 1,64). Cette différence est due principalement au fait que les incomplétudes apparaissent, par définition, à la fin de l'énoncé alors que les faux-départs tendent à être observés en début de l'énoncé.

Les principales constructions observées dans notre corpus ainsi que leurs fréquences sont données dans le tableau suivant :

Structure	Pourcentage
N(XP) v	26,08
N(XP) v infto	15,21
N(XP) coord	15,21
N(XP) pron	8,69
N(XP) prep	8,69
N(XP) v det	6,52
N(XP) det	6,52
N(XP) adv	4,34
N(XP) name	4,34
N(XP) conjonction	2,17
N(XP) v adv	2,17

Tableau 8. Les constructions des incomplétudes observées dans notre corpus et leurs fréquences

La première observation que nous pouvons faire est que, tout comme dans les faux-départs, les constructions verbales sont dominantes dans les incomplétudes. En effet, 50% des cas que nous avons observés se terminent par une construction verbale incomplète. Cette domination est cependant moins claire qu'avec les faux-départs et nous observons une augmentation nette des fréquences d'autres constructions en particulier celle de la coordination. Pour affiner notre analyse, nous avons jugé bon de distinguer entre deux types de coordinations :

1. **Coordination syntaxique :** il s'agit généralement de la coordination de deux arguments du prédicat verbal. Comme il s'agit d'incomplétude, les cas qui sont les coordinations entre les objets selon le schéma suivant : Prédicat_verbal $objet_1$ conjonction_de_coordination $objet_2$.
2. **Coordination discursive :** il s'agit de la coordination qui établit un lien entre deux propositions. Le schéma général de ce genre de coordination est le suivant : $proposition_1$ conjonction_de_coordination $proposition_2$.

Ainsi, après avoir distingué entre les deux formes de coordination, nous avons trouvé que 77,77% des coordinations sont des coordinations discursives. Pour le reste (22,22% des cas), il n'était pas possible pour nous de savoir s'il s'agit d'une coordination syntaxique ou discursive : les indices linguistiques et contextuels n'étaient pas suffisants pour juger.

2.1.1.39 Annotation des fausses extragrammaticalités

Outre les extragrammaticalités, nous avons annoté aussi les énoncés qui contiennent des segments qui ont la forme d'une extragrammaticalité sans en être une. Prenons l'exemple suivant :

That's gonna take the longe(st)- well it's gonna take <sil> **two four six hours** to get back to Corning with those two boxcars (utt32, d93-19.4) (52)

Dans cet exemple, le segment *two four six hours* a la forme d'une autocorrection mais en réalité il s'agit du comptage à haute voix du nombre d'heure que dure le voyage.

L'objectif principal de cette annotation est d'observer les cas d'ambiguïté et de proposer des solutions adaptées à ce problème. Nous avons annoté 159 occurrences de ce genre dans notre corpus.

1.3.4 Les occurrences multiples d'extragrammaticalités

Dans certains contextes, le locuteur peut produire plus d'une extragrammaticalités dans le même énoncé. Selon la relation entre ces extragrammaticalités, nous pouvons distinguer entre deux cas :

2.1.1.40 Les extragrammaticalités multiples

Dans ce cas, l'énoncé contient plusieurs extragrammaticalités complètement séparées, comme dans l'exemple suivant :

Now the problem **is is** that one engine can <sil> pull at most **three three** loaded boxcars (utt55, d93-12.4) (**53**)

Dans cet exemple, nous remarquons que les deux répétitions de *is* et de *three* sont complètement indépendantes l'une de l'autre malgré leur occurrence dans le même énoncé.

Les occurrences multiples peuvent porter sur des phénomènes du même genre (comme dans l'exemple précédent) ou peuvent impliquer deux formes différentes d'extragrammaticalité : une répétition avec une incomplétude, par exemple.

Dans notre corpus, 9,34 des occurrences des extragrammaticalités se trouvent dans une configuration multiple.

2.1.1.41 Les extragrammaticalités imbriquées

Il s'agit de deux extragrammaticalités qui ont, au moins, un mot en commun. Pour mettre au clair cette définition, examinons l'exemple suivant :

(…) <sil> do **I** <sil> **I** need two <sil> do I need two <sil> engines for the (...)

M1 M2 E M22 M3 M4 E M12 M23 M32 M42

(utt39, d92a-1.2)

Figure 27. Exemple de deux extragrammaticalités imbriquées

Dans cet exemple, nous remarquons que la répétition du mot *I* est imbriquée au sein de la répétition de *do I need two*.

D'après nos statistiques, 8,24% des occurrences totales des extragrammaticalités sont des cas *imbriqués*. Comme nous allons le montrer dans le premier chapitre de la quatrième partie, l'imbrication est un phénomène qui nécessite un traitement particulier afin de pouvoir normaliser l'énoncé correctement.

1.3.5 Discussion des résultats de notre annotation

Dans ce paragraphe nous allons discuter les résultats de notre annotation des extragrammaticalités dans le *Trains Corpus* des points de vue linguistique et cognitif. Nous allons en particulier, présenter les deux principales raisons de production des extragrammaticalités que nous avons observées, discuter la régularité des extragrammaticalités que nous avons annoté et finalement parler des implications de nos observations sur un module dédié à la détection et la délimitation des extragrammaticalités.

2.1.1.42 Production des extragrammaticalités

Dans notre analyse du corpus nous avons pu distinguer entre quatre sources principales pour la production des extragrammaticalités :

1. **Non-adéquation sociale :** nous avons observé des cas où les sujets se rendent comptent que la forme linguistique qu'ils ont adoptée ne correspond pas au contexte social du dialogue. Ainsi ils procèdent à une auto-correction pour atteindre un niveau sociolinguistique adapté. Le cas le plus

représentatif de ce genre d'extragrammaticalité est le remplacement des mots oraux et des amalgames par des formes standards comme dans : *I'll uh I will …* ou *yeah yes*. Comme nous pouvons le remarquer dans les deux exemples précédents, les deux formes (remplaçante et remplacée) sont identiques sémantiquement et la seule différence entre elles est le niveau sociolinguistique associé à chacune des formes.

2. **Continuité du message :** l'une des principales raisons de production des extragrammaticalités que nous avons observées est de garder la continuité des messages émis. En d'autres termes, certaines extragrammaticalités ont pour seule fonction de remplir les trous phonétiques dans l'énoncé.
3. **Non-adéquation sémantique :** dans ce cas les extragrammaticalités sont générées pour changer le contenu sémantique du fragment d'énoncé généré.
4. **Non-adéquation linguistique :** les sujets peuvent se rendre compte que la structure qu'ils ont choisie ne permet pas d'établir un lien syntaxique, sémantique et/ou discursif avec les constructions qu'ils ont planifiées de dire après la construction en cours de production. Ainsi, ils effectuent un changement pour pouvoir atteindre leur objectif communicatif.

2.1.1.43 Régularité des extragrammaticalités

Notre annotation du corpus a montré que les extragrammaticalités ne sont pas des phénomènes irréguliers comme on pourrait le penser vue les hétérogénéités des raisons de la production de ces phénomènes. Cela est assez clair avec les répétitions et les autocorrections dont la régularité les rend assez facilement modélisable avec des patrons qui impliquent des connaissances linguistiques assez réduites. En ce qui concerne les faux-départs et les incomplétudes, bien que la régularité de ces phénomènes soit moins évidente à première vue, nous avons constaté dans notre annotation de ces phénomènes que leurs formes semblent être soumises à des considérations grammaticales.

Avant de discuter les cas que nous avons observé, nous allons commencer par une présentation des principes cognitifs clés de la génération du langage oral.

1.3.5.1.1 Principes cognitifs de la génération du langage parlé

Différents travaux dans le domaine de la psycholinguistique expérimentale (Garett, 1988), (Levelt, 1989), ont montré que la conversion d'une forrme conceptuelle pré-verbale en un énoncé parlé est faite selon un nombre de processus (modules) indépendants et spécialisés chacun dans une tâche particulière :

1. Le module de planification sémantique (conceptualiseur) : ce module planifie un contenu sémantique pour être exprimé. Ainsi la sortie de ce module est une représentation sémantique correspondant à ce contenu.
2. Le module de formulation linguistique (le fomaliseur) : ce module effectue la formulation linguistique de la représentation sémantique reçue du conceptualiseur. Cela est fait en sélectionnant les items lexicaux à utiliser et en prenant en considération les contraintes

phonologiques et syntaxiques de la langue utilisée pour la génération. Ainsi, la sortie de ce module est une représentation phonologique et syntaxique abstraite.

3. Générateur de son (articulateur) : la sortie du formulateur est convertie en un signal de parole par le générateur de son (l'appareil articulatoir).

Par ailleurs, il est communément admis que la génération de la parole se fait de manière incrémentale (Kempen et Hoenkamp, 1987), c'est-à-dire, un module ne doit pas attendre la fin du traitement dans le module précédent pour commencer à travailler. Par exemple, le module de génération phonétique peut commencer à générer des sons à partir d'une formulation linguistique du premier fragment de l'énoncé et produit le reste au fur et à mesure de la réception des formulations linguistiques du reste des fragments. Finalement, en ce qui concerne le niveau syntaxique (qui est le sujet de notre discussion), différents travaux ont montré qu'il existe un processus d'amorçage qui permet de planifier les segments syntaxiques à l'avance (Branigan *et al.*, 1995), (Scheepers et Corley, 2000). Ces travaux ont, par ailleurs, montré que le groupe verbal joue un rôle central dans ce processus.

1.3.5.1.2 Génération des répétitions[25]

Les répétitions ont pour fonction de remplir le vide dans l'énoncé afin garder un minimum de continuité dans le message. Selon les principes généraux de la génération de la parole, le mécanisme de génération des répétitions peut être résumé dans les poins suivants :

1. Le module de planification sémantique produit une représentation partielle, cette représentation est formulée linguistiquement et puis générée phonétiquement.
2. Pour des raisons diverses liées au coût cognitif de la tâche ou à l'état psychologique du sujet, le module de planification sémantique tarde à envoyer le segment suivant de la représentation sémantique de l'énoncé à générer.
3. Le module de formulation linguistique décide de répéter le dernier segment généré en attendant la réception de la représentation sémantique du segment suivant.

1.3.5.1.3 Génération des auto-corrections

La génération des auto-corrections se fait selon les deux étapes suivantes :

1. Le module de planification sémantique produit une représentation partielle, cette représentation est formulée linguistiquement et la forme linguistique produite est générée phonétiquement.
2. Le module de planification sémantique se rend compte que la représentation sémantique qu'il vient de produire contient une erreur et décide de reproduire la représentation du segment généré avec la correction de l'erreur.

---

[25] Dans notre discussion de la génération des répétitions et des auto-corrections, nous avons exclu les facteurs biologiques (toux, problèmes de respiration, etc.) étant donné que nous n'avons pas observé des cas de ce genre dans notre corpus.

Comme nous pouvons le remarquer, la production des auto-corrections tout comme celle des répétitions est liée uniquement au dysfonctionnement du module de planification sémantique et n'implique pas le module de formulation syntaxique qui joue un rôle passif dans ce cas.

1.3.5.1.4 Discussion des deux structures syntaxiques les plus fréquemment observées dans les faux-départs et les incomplétudes

Dans notre analyse des faux-départs et des incomplétudes, nous avons observé qu'il existe deux formes dominantes des segments extragrammaticaux : les constructions verbales et les coordinations. Dans ce paragraphe, nous allons discuter ces deux formes à la lumière des principes généraux de la génération du langage parlé que nous avons présenté dans le paragraphe précédant ainsi que les spécificités linguistiques de la langue de notre corpus : l'anglais.

1. **Les constructions verbales :** nous avons vu que les segments extragrammaticaux dans les faux-départs étaient majoritairement de nature verbale (84,49%). De même, la moitié des constructions de ces segments dans les incomplétudes était de nature verbale. Comme nous estimons que cette fréquence est intimement liée à l'ordre canonique des mots en anglais, nous allons commencer par la présentation de celui ci et puis discuter sa pertinence par rapport à la forme des extragrammaticalités produites. En effet, l'anglais est une langue où l'ordre canonique est : SVO (Sujet Verbe Objet). Ainsi, dans cette langue, le prédicat verbal joue un rôle central au sens propre et figuré du terme. Voici une représentation schématique des relations entre le verbe et ses arguments dans les langues SVO :

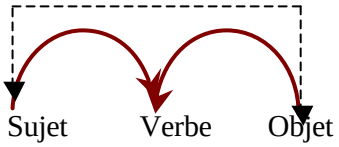


Figure 28. Schéma général des relations entre le prédicat verbal et ses arguments dans les langues SVO

Comme nous pouvons remarquer dans la figure précédente, toute mise en relation du sujet et de l'objet dans les langues SVO nécessite le passage par le prédicat verbal (contrairement aux langues SOV, par exemple, où une première relation entre le sujet et l'objet est établie et dont la nature est clarifiée ultérieurement avec la production du prédicat verbal). Ainsi, nous estimons que la génération des faux départs se fait de la manière suivante :

i. Planification d'un contenu sémantique et génération de la représentation sémantique correspondant à ce contenu (la représentation sémantique peut être formulée et transmise au module suivant de manière incrémentale).

ii. Formulation linguistique incrémentale correspondant à la représentation linguistique obtenue. Cette formulation se fait en construisant la représentation phonologique et syntaxique du premier constituant de l'énoncé à générer et en amorçant celle du (ou des) constituant(s) qui dépendent directement du constituant formulé ou des quels il dépend

directement. Ainsi, dans les langues SVO, le premier constituant formulé est le constituant nominal sujet qui permet d'amorcer uniquement le constituant verbal. Mais, comme nous avons vu au début de ce paragraphe, cela ne suffit pas de juger complètement l'adéquation de la forme partiellement générée par rapport à la représentation sémantique (qui peut être reçue partiellement à ce stade de la génération). Dans la deuxième étape, le prédicat verbal est formulé et le constituant nominal est amorcé. Cela permet de constituer une première formulation linguistique complète de l'énoncé et, par conséquent, juger son degré de correspondance avec la représentation sémantique reçue. Ainsi, lorsque la forme générée est jugée comme étant non-appropriée par rapport au contenu sémantique planifié, cette forme est négligée et une nouvelle formulation est commencée.

2. **Les coordinations :** nous avons vu que les coordinations étaient assez fréquemment observées comme point d'interruption de l'énoncé en particulier dans les incomplétudes (où ils couvrent plus de 15% des cas). Nous avons vu qu'au moins 77,77% des coordinations observées étaient des coordinations discursives. Ainsi, à la lumière des principes cognitifs de génération des extragrammaticalités nous pouvons expliquer ces observations par les points suivants :

   i- Les sujets génèrent le dernier fragment de leur proposition mais ne sont pas encore certains que le contenu sémantique qu'ils veulent exprimer est totalement formulé dans la proposition qu'ils viennent de produire : le module de planification sémantique n'a pas encore donné le signe de fin de représentation sémantique.

   ii- A cause du retard du signe de la fin, le module de formulation linguistique décide qu'une nouvelle proposition est en cours de planification au niveau sémantique et génère la conjonction de coordinations (sans référer à un contenu sémantique explicite de la part du module de planification sémantique) pour lier la proposition produite à la proposition attendue.

   iii- Le module de planification sémantique envoie un signe de fin plutôt qu'une représentation sémantique et l'énoncé produit est incomplet.

1.3.5.1.5 Effet de nos observations sur la génération des extragrammaticalité sur leur analyse

Nos observations sur la génération ainsi que sur la structure des extragrammaticalités ont plusieurs implications par rapport à un module d'analyse des extragrammaticalités :

1. Les extragrammaticalités ne sont pas des phénomènes irréguliers comme on pourrait le penser. Cependant le degré de régularité de ces phénomènes varie d'un phénomène à l'autre (les faux départs sont moins réguliers que les répétitions par exemple). Ainsi, nous pouvons utiliser différentes techniques pour traiter ces phénomènes selon leur degré de complexité.
2. Les connaissances linguistiques ont rôle minimal dans la production des répétitions et des auto-corrections.

3. Les connaissances syntaxiques permettent non seulement de délimiter l'étendue des faux-départs et des incomplétudes mais aussi détecter leur présence (à notre connaissance, tous les travaux précédents ont utilisé la syntaxe pour la délimitation seulement). Pour ce faire, les dépendances syntaxiques des syntagmes au sein de l'énoncé oral doivent être modélisées correctement. En effet, nous avons vu que la non-prise en considération des dépendances syntaxiques peuvent mener à des erreurs de détection ou à des surgénérations.

2 **Chapitre II.2 :** Les formalismes S-TSG et Sm-TAG pour l'analyse grammaticale du langage oral spontané

2.1 Introduction

Depuis le début des études linguistiques, la langue a toujours été considérée comme un niveau de connaissance à part entière bien distinct des autres niveaux de connaissance nécessaires pour l'établissement d'un dialogue : connaissances métalinguistiques, connaissance sur le monde, etc. En effet, cette distinction nette est motivée, sur le plan théorique, par la volonté de la linguistique, qui est une discipline relativement jeune, de s'affirmer comme une branche complètement indépendante de l'investigation scientifique. Sur le plan pratique, cette séparation peut être motivée par le fait que des études interdisciplinaires sont plus difficiles à mener que des recherches mono-disciplinaires étant donné qu'elles nécessitent des connaissances approfondies dans des domaines assez variés. De plus, l'établissement d'un modèle formel *universel* capable de prendre en considération les différents niveaux de connaissances et leurs interactions semble une tâche très difficile dans le contexte de l'état actuel de l'art dans le domaine des sciences cognitives.

Cependant, cette séparation dans le contexte des recherches sur les dialogues orientés vers la tâche ne nous semble pas justifiée. En effet, dans ce contexte, les connaissances sur le monde ainsi que les connaissances linguistiques peuvent être modélisées avec un degré raisonnable de finesse. Cela permet d'explorer de nouveaux modèles qui permettent de rendre compte de l'interaction des différents niveaux de connaissance.

Par ailleurs, comme nous avons vu dans la première partie de cette thèse, près d'un siècle après *la révolution* Saussurienne, dont l'une des principales réalisations est la séparation entre la langue et la parole, l'oral reste un thème marginal dans les travaux dans les domaines de la syntaxe et de la sémantique. En effet, les différentes théories linguistiques sont consacrées à la représentation de l'écrit et négligent presque totalement l'oral qui est pourtant la forme de communication la plus spontanée et la plus courante entre les humains.

Ainsi, nous proposons la Grammaire Sémantique d'Association d'Arbres Sm-TAG comme un formalisme qui tente de combler ce vide dans les travaux précédents. Les propriétés principales de notre formalisme sont :

- La prise en considération des connaissances sur le monde dans la représentation syntaxique des dialogues oraux.

- La prise en considération des phénomènes linguistiques de l'oral dans la définition du formalisme.

Avant de présenter les différentes propriétés de ce formalisme, nous allons commencer par la présentation des éléments syntaxiques de base nécessaires pour la représentation de l'oral. Nous allons ensuite présenter une formalisation de la grammaire sémantique classique qui était notre premier pas pour la proposition de notre formalisme.

2.2 Les éléments de base pour une théorie syntaxique et leur pertinence pour la représentation de l'oral

Les connaissances syntaxiques peuvent être divisées en deux sources principales :

2.2.1 Le système casuel

Il s'agit de l'ensemble des moyens utilisés par une langue pour marquer les rôles syntaxiques (sujet, objet, etc.). En français, ces moyens sont :

a. **La topologie :** il s'agit de l'ordre selon lequel les mots sont agencés au sein de la phrase. En général, la topologie permet de savoir la fonction d'un argument selon sa position par rapport au verbe (Lazard, 1994). Par exemple, le français est une langue à ordre SVO (Sujet Verbe Objet). Selon les langues, cet ordre peut varier entre fixe et totalement variable. Comme nous avons vu dans la première partie de cette thèse, le français oral tend à être une langue à ordre fixe.

b. **Les prépositions :** les prépositions indiquent le cas du syntagme qui vient après (vocatif, datif, etc.).

c. **Flexion casuelle :** en français, ce moyen est limité à la distinction entre pronoms *je* (sujet) et *me* (objet).

2.2.2 Accord en genre et en nombre

En français, il s'agit d'un mécanisme selon lequel un nom ou un pronom donné exerce une contrainte formelle sur les pronoms qui le représentent, sur les verbes dont il est sujet, sur les adjectifs ou participes passés qui se rapportent à lui (Dubois, 1994). L'accord est généralement utilisé pour résoudre certaines ambiguïtés d'attachement surtout en cas de dépendances lointaines (qui ne sont pas très fréquentes à l'oral).

2.2.3 Quelles sources d'informations pour le traitement du français oral ?

Pour sélectionner les sources d'informations à considérer dans la Sm-TAG, deux critères ont été retenus :

1. **L'intérêt de l'information :** elle est jugée selon l'importance du rôle que joue cette information dans le traitement de l'énoncé.
2. **La fiabilité de l'information :** elle est jugée selon la régularité de celle-ci ainsi que la possibilité de son bruitage.

Ainsi, nous avons retenu le système casuel dans notre formalisme. En fait, d'une part, il s'agit à la fois d'une information nécessaire et fiable pour le traitement. Nécessaire, puisqu'elle permet de préciser les différents rôles syntaxiques et sémantiques. Elle est fiable à cause de la régularité des différents moyens de marquage casuel en français.

Par ailleurs, l'accord n'a pas été considéré dans la Sm-TAG. D'une part, l'information qu'il véhicule n'est pas centrale dans le traitement et d'autre part, il s'agit d'une information non fiable. En fait, les erreurs d'accords sont parmi les erreurs les plus fréquentes des systèmes de reconnaissance de la parole sans oublier les cas assez fréquents de non-respect de l'accord en français oral dont les expressions clivées constituent l'exemple typique : *c'**est** des truc**s*** (au lieu de *ce **sont** des trucs*).

2.3 La grammaire sémantique de substitution d'arbres (S-TSG)[26]

La S-TSG est une formalisation que nous avons proposée de la grammaire sémantique classique. Notre présentation sera limitée aux aspects formels étant donné que les différentes propriétés de la grammaire sémantique ont été présentées dans la première partie de cette thèse.

2.3.1 Les unités de base dans la S-TSG

Les arbres constituent les unités de base dans la S-TSG. Contrairement à la LTAG, ces arbres ne sont pas forcément ancrés par un item lexical. Par ailleurs, il n'existe pas d'arbres auxiliaires dans la S-TSG, les arbres initiaux étant les seuls arbres possibles dans ce formalisme. La profondeur des arbres de tous types est limitée à une branche. Cela veut dire que nous pouvons représenter ce formalisme à la fois de manière syntagmatique ou comme un formalisme d'arbres (Abeillé, 1993). Dans ce travail, nous avons opté pour la représentation comme un formalisme d'arbres principalement pour pouvoir le comparer à l'autre formalisme que nous avons proposé (la Sm-TAG) ainsi qu'à d'autres formalismes comme LTAG et ses dérivés.

2.1.1.44 Les arbres lexicaux

Il s'agit d'arbres dont la racine correspond à une catégorie sémantique associée à un item lexical qui est l'ancre de cet arbre. En voici quelques exemples d'arbres lexicaux :

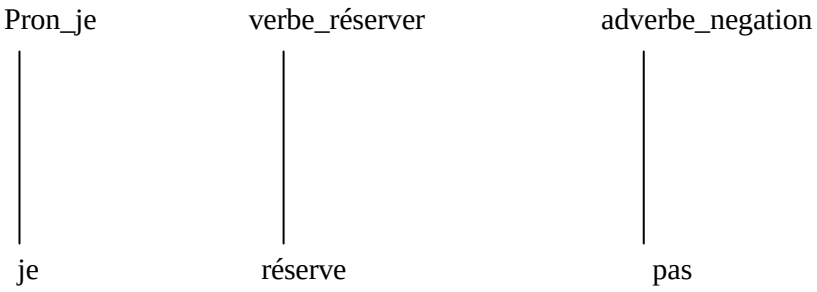


Figure 29. Exemples d'arbres lexicaux dans la S-TSG

Les arbres servent à lier le lexique à des structures supérieures qui sont les arbres locaux.

[26] S-TSG est l'acronyme de : Semantic Tree Substitution Grammar.

2.1.1.45 Les arbres locaux

Les arbres locaux correspondent aux segments conceptuels dans la grammaire sémantique. Comme nous avons vu dans la première partie de cette thèse, ces unités ne sont pas définies selon des critères clairement définis. Voici quelques exemples de ces arbres :

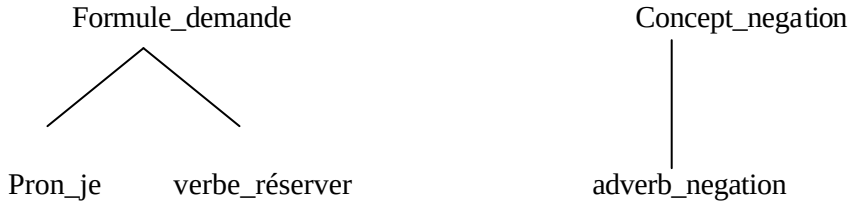


Figure 30. Des arbres locaux dans le formalisme S-TSG

Nous remarquons que les catégories dans ces arbres sont toutes de nature sémantique ou syntaxico-sémantique comme la catégorie (pron_je).

2.1.1.46 Les arbres globaux

Les arbres globaux sont destinés à lier les arbres locaux en unités plus importantes et représenter leurs dépendances sémantiques servant ainsi à les désambiguïser. Voici quelques exemples d'arbres globaux :

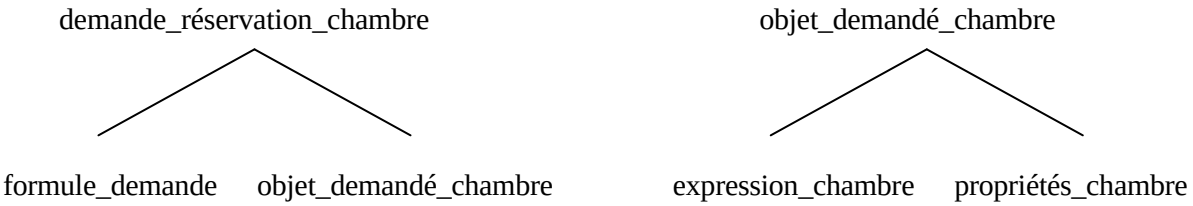


Figure 31. Exemple de deux arbres globaux dans le formalisme S-TSG

Comme nous pouvons le voir dans l'exemple précédent, un arbre global peut dominer un arbre local (formule_demande) ou un autre arbre global dans certains cas (objet_demandé_chambre) mais pas d'arbres lexicaux directement.

2.3.2 L'opération de combinaison

En S-TSG, seule l'opération de substitution (similaire à celle du formalisme LTAG) est utilisée pour combiner les arbres. L'adjonction n'est pas possible dans ce formalisme. Voici un exemple de substitution :

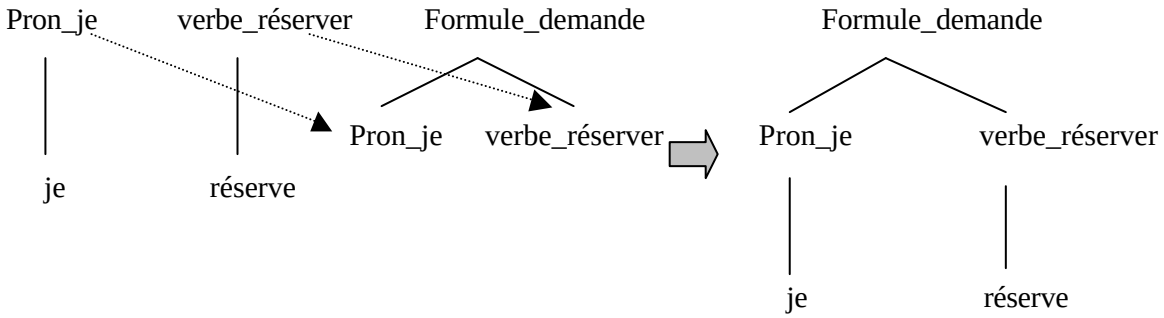


Figure 32. Un exemple de substitution dans le formalisme S-TSG

2.3.3 Définition formelle de la S-TSG et son équivalence avec une CFG

D'un point de vue formel, la S-TSG peut être définie comme un quadruplet (Σ, NT, I, S) où :

- S est un ensemble fini de symboles terminaux.
- **NT** est un ensemble de symboles non-terminaux. Les symboles non-terminaux sont des catégories sémantiques dérivées d'un modèle de la tâche du dialogue.
- **I** est un ensemble fini d'arbres élémentaires appelés arbres initiaux. La profondeur de ces arbres est limitée à une branche. Les nœuds internes ainsi que les nœuds sur les frontières peuvent être annotés avec des symboles terminaux ou non-terminaux. Les non-terminaux du nœud frontière sont marqués pour la substitution. Contrairement aux grammaires classiques les non-terminaux sont généralement de nature sémantique.
- **S** est un symbole non-terminal distingué (S est l'axiome de la grammaire). Contrairement aux approches classiques, l'axiome d'une grammaire peut être une multitude de symboles. Ainsi, dans certains cas, tous les non-terminaux de la grammaire peuvent être l'axiome de la grammaire. Cette différence, permet de faire des analyses partielles.

La S-TSG est un formalisme fortement équivalent à une CFG. Pour prouver cette équivalence, il faut prouver le théorème suivant : pour toute grammaire S-TSG $G = (S, NT, I, S)$ il existe une CFG $G' = (S, NT, P, S)$ qui génère le même langage.

La preuve de ce théorème est un processus trivial (voir une preuve similaire dans (Shabes et Waters, 1995)), il faut remplacer tous les arbres élémentaires *t* par des règles de réécriture *R*. Pour ce faire, il faut suivre les démarches suivantes : l'étiquette de la racine de *t* devient la partie gauche de la règle *R*. Les étiquettes sur la frontière de *t* deviennent la partie droite de *R*.

Par ailleurs, vu que la profondeur des arbres de la S-TSG est limitée à une branche, l'arbre de dérivation ainsi que l'arbre dérivé sont identiques tout comme dans les CFGs.

2.3.4 Portée et limites de la S-TSG

Les avantages de la S-TSG peuvent être résumés dans les deux points suivants :

1. **Théorique :** étant défini formellement et linguistiquement, la S-TSG rend possible la comparaison de la grammaire sémantique avec les autres formalismes et permet d'établir des bilans pour juger l'adaptation de ce formalisme par rapport à une tâche particulière comparé à d'autres formalismes candidats à l'utilisation pour cette tâche.
2. **Pratique :** l'avantage principal de la S-TSG par rapport à la grammaire sémantique classique est la distinction entre les trois types d'arbres : les arbres lexicaux, locaux et globaux. Cela facilite la tâche d'écriture de la grammaire ainsi que de sa modification. Par ailleurs, cela rend l'enseignement de ce formalisme plus facile.

N'étant qu'une formalisation de la grammaire sémantique, la S-TSG présente tous les inconvénients de ce dernier : pauvreté syntaxique, non-pertinence linguistique, etc. Ces inconvénients théoriques, ont été confirmé après notre implantation et notre test d'une grammaire S-TSG au sein de notre système SAFIR que nous allons présenter en détail plus loin dans la quatrième partie de cette thèse. Ainsi, nous avons proposé une version avancée de ce formalisme qui combine les avantages de la grammaire sémantique à ceux des grammaires syntaxiques classiques. Nous avons baptisé ce formalisme la grammaire sémantique d'association d'arbres Sm-TAG.

2.4 La Grammaire Sémantique d'Association d'Arbres (Sm-TAG)

La Sm-TAG est un formalisme hybride (syntaxique / sémantique[27]) basé sur l'unification. La propriété essentielle de la Sm-TAG est de permettre une linéairisation directe des structures sémantiques fonctionnelles à celles des structures syntaxiques. Ainsi, nous avons un seul arbre pour représenter la phrase au lieu d'un arbre séparé pour la syntaxe et un autre arbre pour la sémantique comme nous avons vu avec les TAGs synchrones par exemple. Nous avons proposé la Sm-TAG comme un compromis entre, d'une part, les grammaires syntaxiques classiques qui ne permettent pas d'obtenir une analyse robuste et la grammaire sémantique qui ne fournit pas une analyse profonde.

Bien que la Sm-TAG présente des propriétés intéressantes pour d'autres tâches comme la génération dans le contexte de systèmes de dialogues oraux spontanés, nous allons nous concentrer dans notre présentation et argumentation sur ses avantages pour l'analyse du langage oral spontané étant donné que cette application constitue l'objectif principal de notre thèse.

2.4.1 Définition fonctionnelle de la Sm-TAG

D'un point de vue fonctionnel, un formalisme comme la Sm-TAG peut être défini selon trois facteurs :

2.1.1.47 La sortie de la grammaire

La sortie de la grammaire est une représentation logique correspondant à l'analyse de la phrase. Cette représentation a la forme d'un ensemble d'arbres annotés avec des labels correspondant aux différentes catégories syntaxiques et sémantiques.

[27] Le mot sémantique est utilisé ici au sens large du terme.

2.1.1.48 Les unités de base

Les arbres élémentaires sont divisés en trois parties : des arbres lexicaux, des arbres locaux et des arbres globaux. Cette division est basée sur des critères syntaxiques, sémantiques et pragmatiques.

1. **Les arbres lexicaux :** les arbres lexicaux sont les unités les plus simples dans la Sm-TAG. Ils constituent le noyau à la fois syntaxique et lexical du formalisme. Il s'agit généralement d'arbres dont la racine est étiquetée par une catégorie syntaxique et qui sont ancrés chacun par un item lexical. Deux types d'arbres lexicaux sont utilisés :

   i- **Les arbres lexicaux auxiliaires :** il s'agit d'arbres de profondeur 2 correspondant aux modifieurs (adverbes, adjectifs, etc.) et qui se lient aux autres arbres par l'opération d'association (cf. section 5.1.1.3.).

   ii- **Les arbres lexicaux initiaux :** il s'agit d'arbres de profondeur 1 ou 2 correspondants aux items lexicaux normaux et qui s'associent aux autres arbres par l'opération de substitution (cf. section 5.1.1.3.).

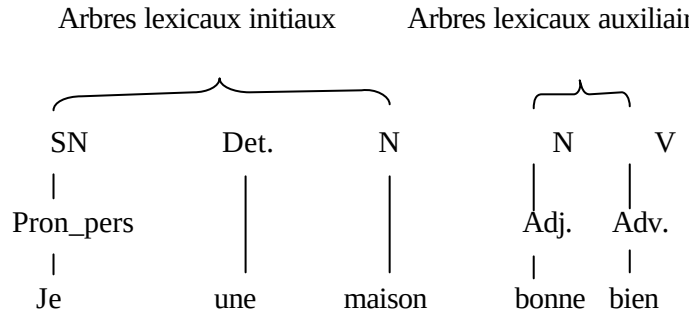


Figure 33. Exemples d'arbres lexicaux

Voici un tableau général qui représente les propriétés clés de notre formalisme :

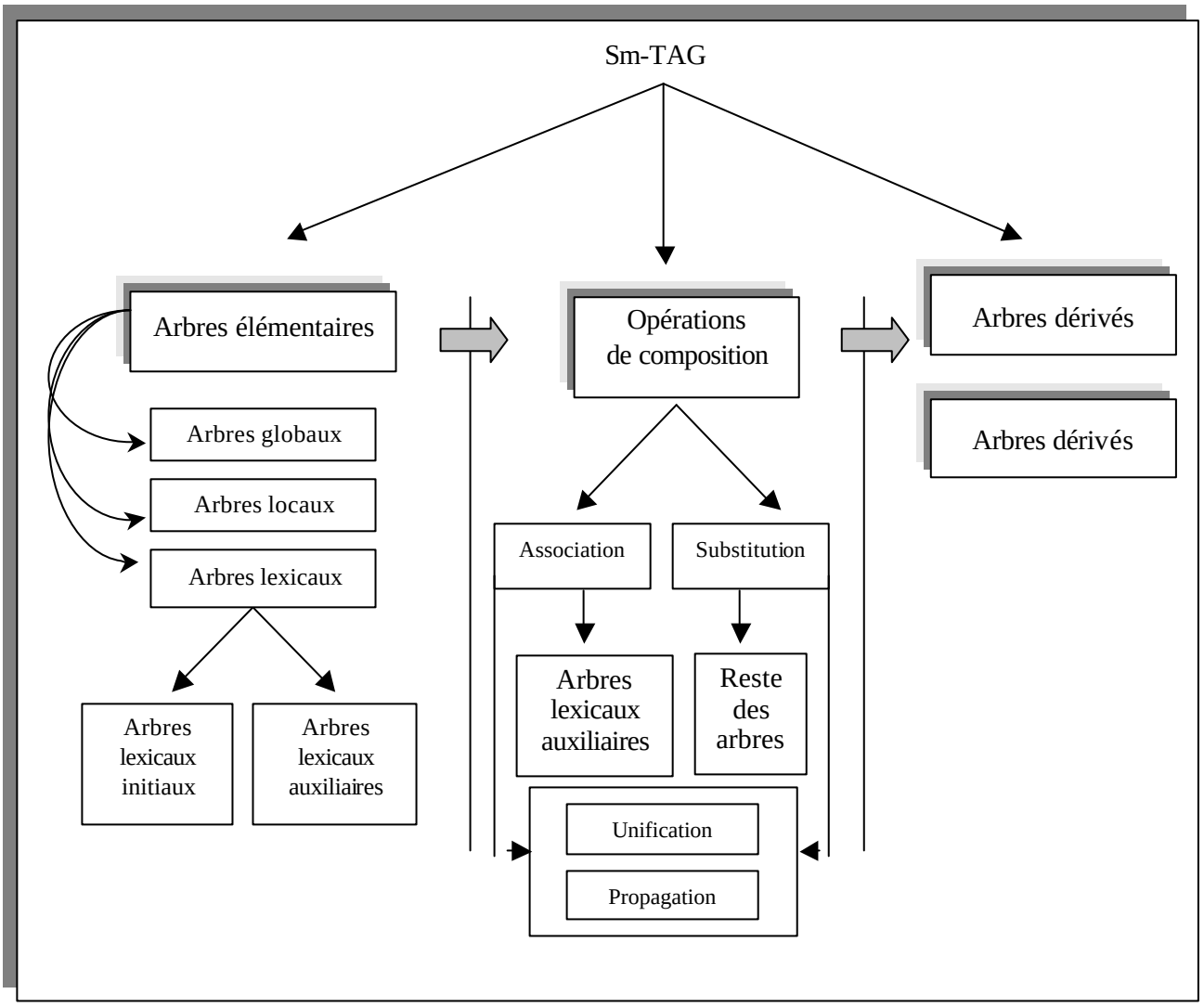


Figure 34. Schéma fonctionnel de la Sm-TAG

2. **Les arbres locaux :** il s'agit d'un ensemble d'arbres dont les racines sont annotées avec des symboles non-terminaux et dont les nœuds feuilles sont annotés avec des non-terminaux ou avec des terminaux. La profondeur maximale de ces arbres est limitée à une seule branche. Les arbres locaux sont construits selon trois principes de bonne formation :

   i. **Principe de consistance sémantique :** chaque arbre local doit avoir une représentation sémantique non vide.

ii. **Principe de non compositionnalité sémantique :** chaque arbre local correspond à une unité sémantique unique. Une unité sémantique est définie selon un ensemble de considérations sémantiques et communicatives dont les principales sont[28] :

a- **Topicalité :** dans l'énoncé certains segments jouent le rôle de thème qui indique ce dont parle le locuteur. D'autres segments peuvent jouer le rôle du rhème. Le rhème est un segment qui donne des informations portant sur le thème. Contrairement aux approches classiques, ce que nous considérons comme thème ou rhème n'est pas forcément le thème ou le rhème global de l'énoncé mais nous concevons plutôt la relation thème-rhème à un niveau local qui marque la relation de détermination sémantique des segments les uns par rapport aux autres.

b- **Donné *vs.* non donné :** on distingue ce que le système connaît *a priori* (par le modèle de la tâche) de ce qui est nouveau.

c- **Importance :** on distingue ce qui est souligné comme important de ce qui est secondaire. Dans la Sm-TAG ce critère a une valeur binaire. C'est-à-dire on distingue uniquement entre deux types d'unités :

- Des unités pertinentes qui sont considérée comme arbres élémentaires dans la grammaire.
- Des unités non pertinentes qui ne sont pas considérées comme arbres élémentaires puisque l'information qu'elles véhiculent n'est pas nécessaire pour la tâche du système. Par exemple, dans l'énoncé :

  allô oui c'est le bureau du ministre j'aimerais avoir des informations sur la disponibilité de votre suite…

  Dans cet énoncé, le segment *c'est le bureau du ministre,* n'est pas considéré comme pertinent dans le contexte d'un système de réservation automatique de chambres puisque la fonction du client n'étant pas considérée comme un critère qui exige une réaction particulière de la part du système et par conséquent il n'est pas associé à un arbre élémentaire.

Ce critère est à la base de la stratégie sélective qui permet de localiser les segments pertinents dans le message.

Pour rendre les principes de segmentation plus concrets, examinons l'exemple suivant : *je voudrais réserver un billet de train*.

[28] Voir (Andrews, 1985) pour la présentation de principes similaires dans le contexte de la syntaxe typologique et fonctionnelle.

- Le mot *je* ne peut pas constituer un segment puisqu'il est donné (on sait à priori que l'interlocuteur est un client).
- Les segments *je voudrais* et réserver peuvent constituer des arbres élémentaires puisqu'ils constituent une articulation thématique (ou une relation thème/rhème) qui véhicule une information importante pour la tâche.
- Les mots *un* et *billet* ainsi que les mots *de* et *train* ne peuvent pas constituer des segments indépendants puisqu'ils ne font pas partie d'une articulation thématique.
- Les segments *un billet* et *de train* constituent dans le contexte d'un dialogue multi-domaine (dans lequel on peut avoir une demande de billet d'avion par exemple) une articulation thématique informative. Par contre, dans le contexte d'un dialogue de réservation de billets de trains uniquement, un billet de train constitue un seul segment, puisque l'articulation entre *un billet* et *de train* n'est pas informative.

Ainsi, le résultat de la segmentation dans le contexte d'un dialogue multi-domaine est le suivant : [je voudrais] [réserver] [un billet] [de train]. Dans le contexte d'un système de réservation de billets de trains uniquement la segmentation est la suivante : [je voudrais] [réserver] [un billet de train].

iii. Principes syntaxiques : contrairement à LTAG, la construction des arbres locaux est essentiellement basée sur la sémantique. Cependant, la syntaxe n'est pas totalement exclue de la segmentation. Ainsi, en Sm-TAG les principes syntaxiques sont utilisés pour contrôler les principes sémantiques en cas d'ambiguïté ou d'insuffisance de ceux-ci par exemple. En d'autres termes, on peut avoir des arbres locaux qui violent les principes syntaxiques, mais lorsque les principes sémantiques autorisent une multitude de segmentations, la priorité est donnée aux arbres qui respectent les principes syntaxiques.

Le principe de co-occurrence prédicat argument est le principe syntaxique le plus important dans Sm-TAG. Prenons comme exemple l'énoncé : *Oui c'est pour deux personnes*. Cet énoncé peut être segmenté de deux manières selon les critères sémantiques :

[Oui] [c'est] [pour deux personnes]

[Oui][c'est pour] [deux personnes]

Parmi ces deux possibilités, seule la deuxième sera retenue étant donné qu'elle est la seule à satisfaire la condition de co-occurrence du prédicat (le verbe être) et ses arguments (le démonstratif *ce* et la préposition *pour*).

3. Les arbres globaux : il s'agit d'un ensemble d'arbres dont les nœuds racines et feuilles sont annotés avec des non-terminaux. Le rôle de ces arbres consiste à assembler les arbres locaux ou globaux en segments plus importants. La bonne formation d'un arbre global est basée sur le principe de co-occurrence d'un prédicat et de ses arguments. La relation prédicat/argument est,

elle aussi, basée sur des critères essentiellement sémantiques. Des exemples d'arbres élémentaires locaux et globaux sont présentés ci-dessous dans la figure 48 :

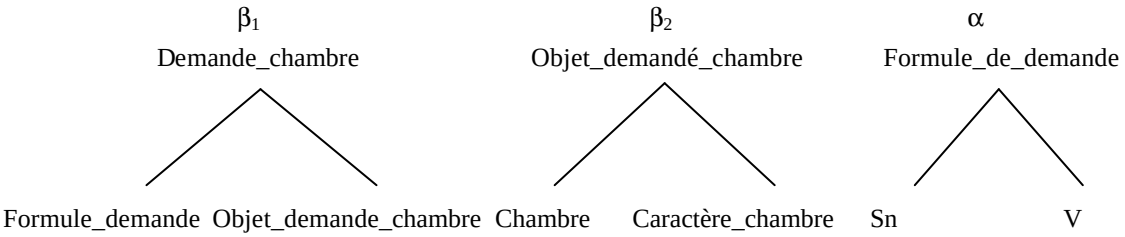


Figure 35. Arbres élémentaires locaux et globaux

2.1.1.49 Les opérations de composition

Il s'agit des opérations qui permettent d'unifier les arbres élémentaires en arbres de dérivation. Deux opérations sont utilisées dans la Sm-TAG :

2.4.1.1.1 L'opération de substitution

L'opération de substitution dans la Sm-TAG est similaire à la substitution dans les formalismes LTAG et S-TSG que nous avons présentés précédemment. Toutes fois, voici un exemple de substitution dans la Sm-TAG :

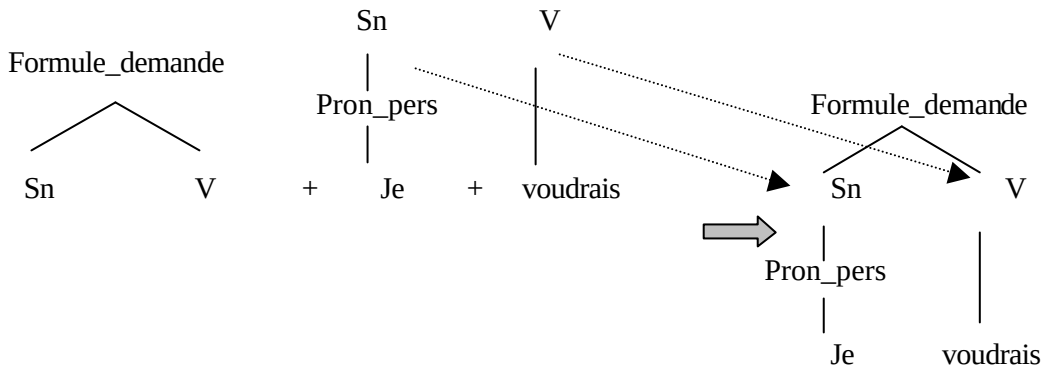


Figure 36. L'opération de substitution

2.4.1.1.2 L'opération d'association

Cette opération est très proche de l'opération d'adjonction du formalisme TIG (Tree Insertion Grammar) (Schabes, 1994) et de l'opération de Furcation du formalisme TFG (Tree-Furcating Grammar), (Cavazza et constant, 1996), (Roussel, 1999).

Les différences principales entre cette opération et l'opération d'adjonction classique des LTAGs se résument dans les points suivants (Schabes, 1994) :

a. Les arbres auxiliaires englobants sont interdits ainsi que les arbres auxiliaires vides. Ce qui conduit à limiter les arbres auxiliaires uniquement aux arbres auxiliaires gauches ou aux arbres auxiliaires droit.

b. Il est interdit qu'un arbre auxiliaire gauche (droit) s'associe à un nœud situé à l'épine dorsale (le chemin entre la racine et le pied de l'arbre) à gauche (droit) de l'arbre auxiliaire.

c. L'association est aussi interdite avec un nœud *h* qui est situé à droit (gauche) de l'épine dorsale de l'arbre auxiliaire gauche(droit) *T*. sachant que, pour qu'un arbre *T* soit un arbre gauche (droit), chaque nœud frontière doit être étiqueté avec ε.

Voici les schémas correspondants à l'adjonction englobante et aux différents types d'association :

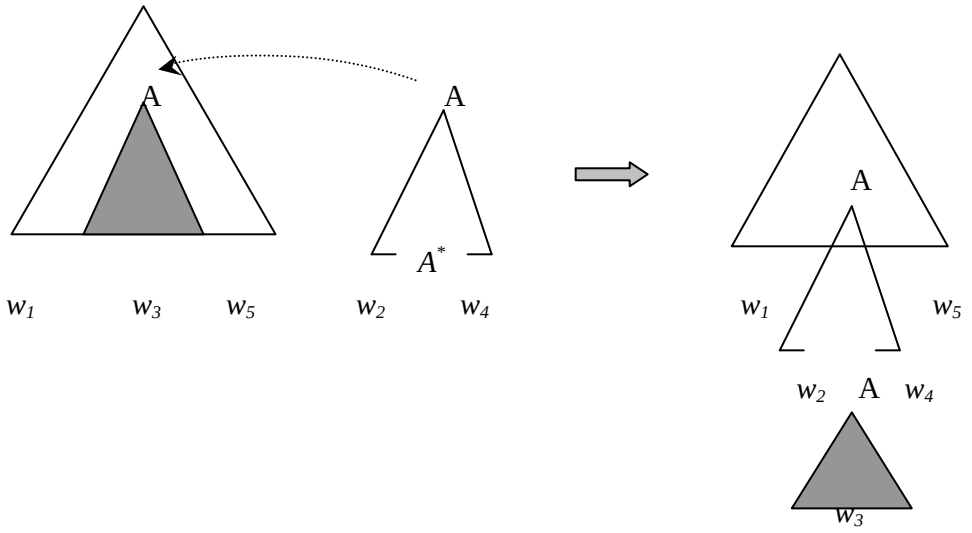


Figure 37. Adjonction englobante interdite en TIG et en Sm-TAG

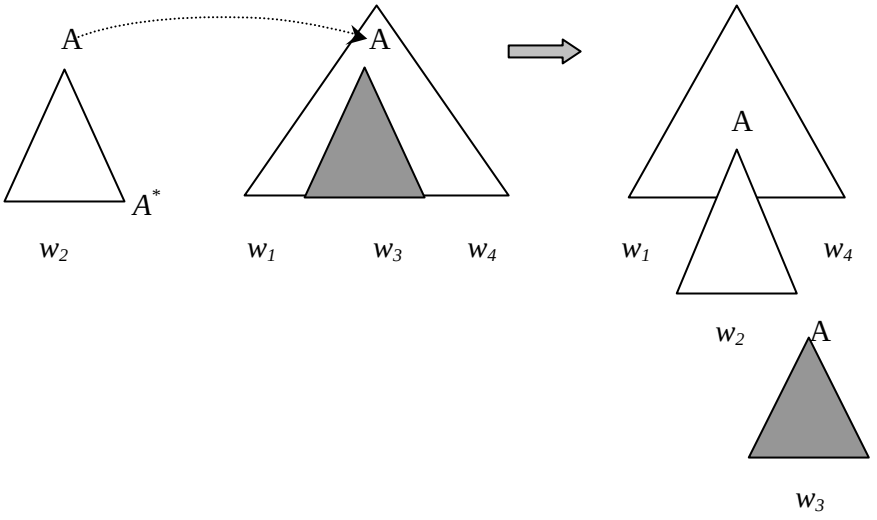


Figure 38. Association gauche

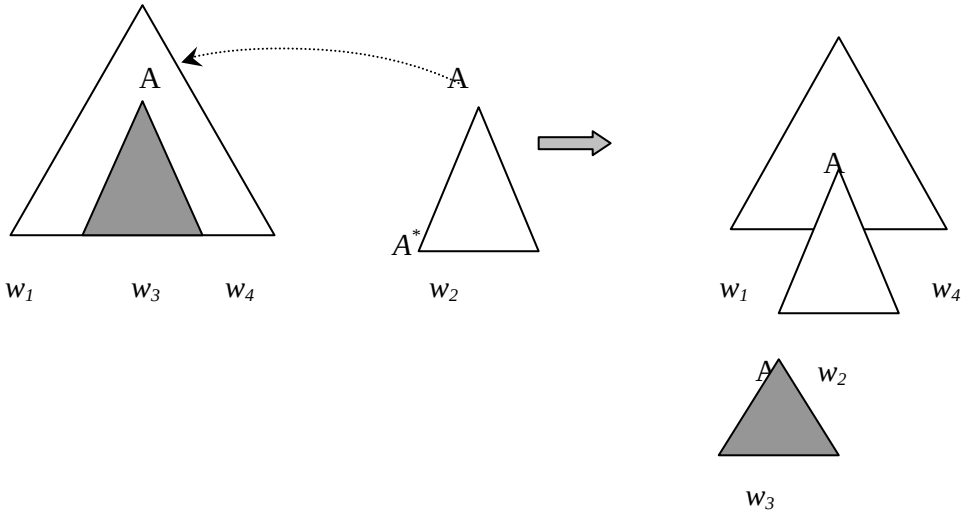


Figure 39. Association droite

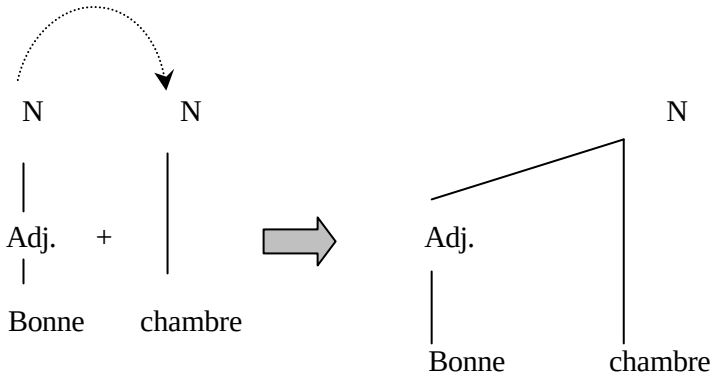


Figure 40. Un exemple d'association

En d'autres termes, la différence principale entre l'opération d'association et l'adjonction classique dans LTAG est que l'opération d'association permet la composition des arbres sans avoir recours à un niveau supplémentaire. Par ailleurs, comparée à l'adjonction de la TIG, l'association peut être vue comme une restriction de cette dernière. En effet, l'adjonction simultanée est abandonnée dans la Sm-TAG puisqu'elle n'est pas d'une utilité réelle pour notre analyse.

L'ajout de l'opération d'association a pour effet d'augmenter la générativité de la grammaire en permettant une intégration souple des modifieurs avec le reste des structures élémentaires de la grammaire.

2.4.2 Définition formelle

Formellement, la Sm-TAG peut être représentée par un quintuplet (Σ, NT, S, I, A) où :

S est un ensemble fini de symboles terminaux.

NT est un ensemble de symboles non-terminaux. Les symboles non-terminaux peuvent être des catégories syntaxiques ou sémantiques.

S est un symbole non-terminal distingué (S est l'axiome de la grammaire). Contrairement aux approches classiques, l'axiome d'une grammaire peut être une multitude de symboles. Ainsi, dans certains cas, tous les non-terminaux de la grammaire peuvent être l'axiome de la grammaire. Cette différence, permet de faire des analyses partielles dans lesquelles des constituants sous-phrastiques peuvent être considérés comme des structures bien formées.

I est un ensemble fini d'arbres élémentaires appelés arbres initiaux. Les nœuds internes ainsi que les nœuds sur les frontières sont annotés avec des symboles terminaux ou non-terminaux. Les non-terminaux du nœud frontière sont marqués pour la substitution.

A les arbres auxiliaires sont les arbres caractérisés par les points suivants :

- Les nœuds internes sont annotés avec des symboles non-termiaux.
- Les nœuds sur les frontières sont annotés avec des symboles terminaux (les ancres des arbres)[29].

2.1.1.50 La dérivation dans Sm-TAG

La composition des arbres dans la Sm-TAG, tout comme LTAG ou n'importe quel autre formalisme à base d'arbres dont la profondeur est supérieure à une branche, peut être représentée de deux manières : avec les arbres dérivés qui représentent le produit de la composition d'une part et d'autre part, les arbres de dérivation qui représentent la manière dont ce produit a été obtenu.

2.1.1.51 L'équivalence avec une CFG

Tout d'abord, tout comme dans la TIG (Schabes et Waters, 1994), toute CFG peut être convertie trivialement en une Sm-TAG qui génère les mêmes arbres. Cela est possible en remplaçant toute règle *R* par un arbre de profondeur 1. Les éléments de la partie droite de la règle *R* deviennent les étiquettes de l'arbre ainsi crée, avec des non-terminaux marqués pour la substitution. Si la partie droite de *R* est vide, l'arbre élémentaire créé a un seul élément de frontière marqué avec *e* Parallèlement, une Sm-TAG qui n'utilise pas d'arbres auxiliaires (et par conséquent n'utilise pas l'opération d'association) et qui contient uniquement des arbres initiaux de profondeur 1, peut être convertie automatiquement en une CFG en remplaçant les arbres initiaux de cette grammaire par des règles de récriture CFG.

---

[29] Tous les arbres auxiliaires sont des arbres lexicaux qui sont, comme leur nom l'indique, ancrés par un item lexical.

Pour prouver formellement l'équivalence générative entre la Sm-TAG et la CFG il faut prouver que pour tout langage généré par une grammaire Sm-TAG G = (Σ, NT, S, L_t) il existe une grammaire CFG G'= (Σ, NT', S, L_t) qui génère le même langage.

Ce théorème a été prouvé pour l'opération d'adjonction de la TIG par (Schabes, 1994). Dans ce qui suit nous adaptons cette preuve pour la Sm-TAG, l'association étant simplement une restriction de l'adjonction de la TIG. L'idée principale de la preuve de ce théorème est basée sur l'élimination des arbres auxiliaires pour arriver à une version avec des arbres initiaux uniquement et dont la conversion en CFG est triviale comme nous avons vu. Les étapes de cette preuve sont les suivantes :

- Pour chaque non-terminal A_i dans *NT*, ajouter deux non-terminaux supplémentaires Y_i et Z_i pour créer un nouvel ensemble de non-terminaux *NT'*.
- Pour chaque non-terminal A_i ajouter les règles suivantes à *P :* $Y_i \rightarrow e$ *et* $Z_i \rightarrow e$.
- Changer tous les nœuds *m* dans chaque arbre élémentaire dans *I et A* de la manière suivante : soit A_i l'étiquette de *m.*. Si et seulement si une association gauche est possible à *m* alors ajouter un fils gauche à *m* étiqueté avec Y_i et le marquer pour la substitution. Si et seulement si une association droite est possible à *m* alors ajouter un nouveau fils droit de *m* étiqueté Z_i et le marquer pour la substitution.
- Convertir tous les arbres auxiliaires *t* dans *A* en arbres initiaux de la manière suivante : soit A_i une étiquette de la racine *m* de *t*. Si *t* est un arbre auxiliaire gauche, alors ajouter une nouvelle racine étiquetée *Yi* avec deux fils : *m* à gauche et à droite un nœud étiqueté *Yi* et marqué pour la substitution. Sinon, ajouter une nouvelle racine étiquetée *Zi* avec deux fils : *m* à gauche et à droite un nœud étiqueté avec *Zi* et marqué pour la substitution. Changer l'étiquette du nœud pied de *t* avec *e.*, ce qui rend ainsi *t* un arbre initial.
- Maintenant, tous les arbres *t* sont des arbres initiaux. Chacun de ces arbres peut être converti en une règle *R* dans *P* de la manière suivante : l'étiquette de racine de *t* devient la partie gauche de *R*. Les étiquettes sur la frontière de *t* avec n'importe quelle occurrence de *e* omis, deviennent la partie droite de *R*.

Il n'est pas inutile de rappeler que *G'* génère uniquement les mêmes chaînes que *G* mais il ne génère pas les mêmes arbres.

2.4.3 Les aspects sémantiques de la Sm-TAG

2.1.1.52 Catégorisation

Les catégories associées aux nœuds des arbres élémentaires peuvent consister tout simplement en catégories syntaxiques classiques ou en catégories sémantico-pragmatiques déduites directement d'une ontologie superficielle de la tâche. L'ontologie contient en général les concepts clés de l'application ainsi que les relations de dépendances entre eux.

2.1.1.53 Représentation des traits

Les nœuds des arbres élémentaires sont *décorés* d'un ensemble de traits de natures diverses. Ces traits servent à contraindre l'unification des nœuds des arbres selon des critères syntaxiques et sémantiques. Les traits peuvent être des traits syntaxiques classiques (nombre, genre, etc.) ainsi que des macro-traits MTs induits directement de l'ontologie la tâche. Ces MTs constituent la différence principale entre la Sm-TAG et les grammaires d'unification classiques. Le choix du type de traits correspondant à un nœud est essentiellement dépendant de la fréquence et de la fonction de ce nœud. Les traits syntaxiques sont utilisés pour les items lexicaux partagés entre les différents arbres. Cela permet de faire le partage des ressources lexicales entre les arbres (ce qui n'est pas le cas avec les grammaires sémantiques classiques). Les MTs sont généralement utilisés pour les items propres à chaque arbre ce qui permet de faire l'économie de la vérification d'un ensemble de traits syntaxiques redondants. Par ailleurs, les traits sont divisés en deux parties (comme dans les LTAGs). Nous avons, d'une part, des traits amonts qui indiquent la relation d'un nœud avec les nœuds qui le dominent, et d'autre part, nous avons des traits avals qui indiquent la relation du nœud avec ceux qu'il domine. Voici quelques structures de traits :

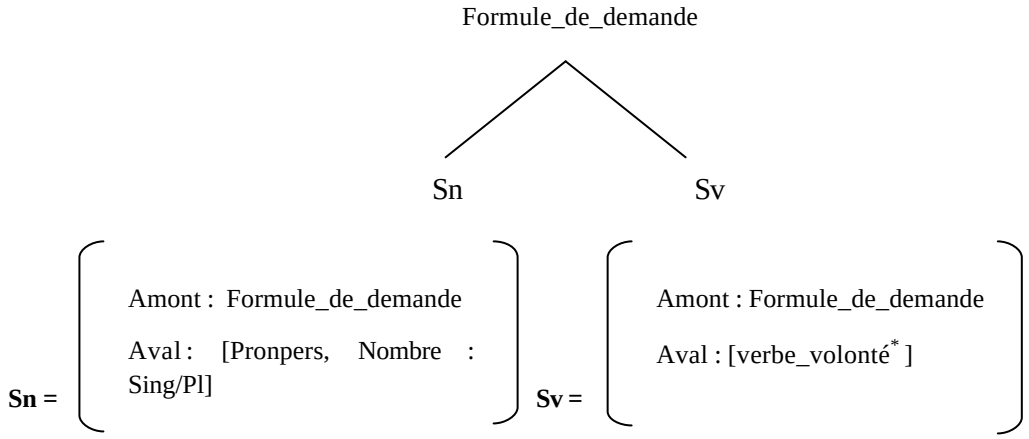


Figure 41. Exemple de structures de traits possibles dans le formalisme Sm-TAG

L'existence des traits ainsi que leur finesse est un paramètre que l'on peut modifier selon les besoins et les moyens. Ainsi, une grammaire Sm-TAG peut être écrite sans traits, uniquement avec des traits syntaxiques et sémantiques classiques ou uniquement avec des macro-traits ou bien comme dans l'exemple présenté ci-dessus avec des traits hybrides.

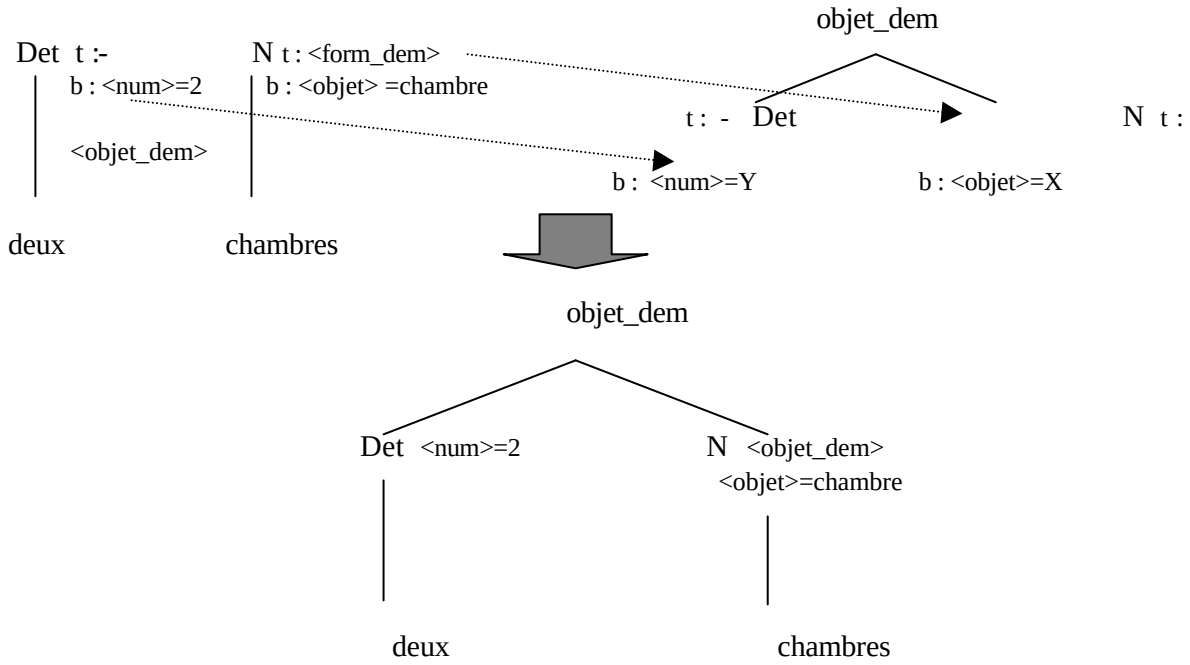
2.1.1.54 Unification et propagation sémantique

Comme nous venons de voir dans les deux paragraphes précédents, l'information sémantique est représentée de deux manières au sein des arbres élémentaires. Au niveau des arbres lexicaux, la sémantique est représentée sous formes de traits qui enrichissent les non-terminaux de la grammaire alors qu'elle est codée directement dans les non-terminaux des arbres locaux et globaux. Ainsi, nous avons deux opérations qui correspondent à ces deux types d'informations :

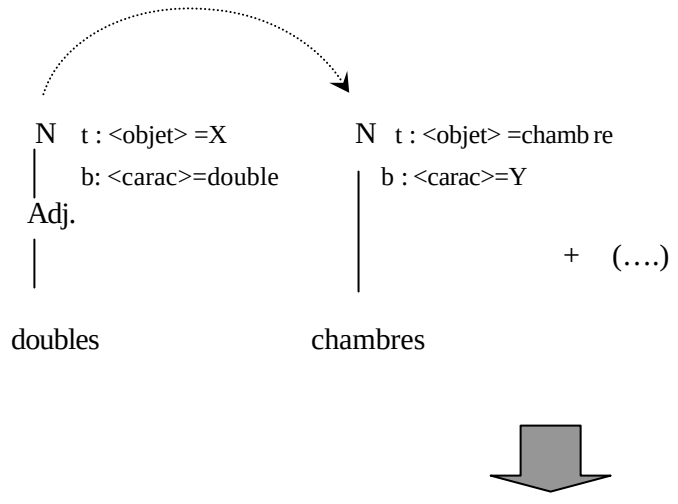
2.4.3.1.1 L'unification

Cette opération a une double fonction : d'une part elle formule des contraintes sur les deux opérations syntaxiques (la substitution et l'association) et d'autres elle gère la propagation de ces contraintes au cours de l'analyse. Elle consiste à vérifier que les traits amonts d'un nœud racine de l'arbre substitué s'unifient avec les traits amonts du nœud où a lieu la substitution. Dans le cas d'une association, il doit y avoir, d'une part, unification du trait amont de la racine de l'arbre auxiliaire avec les traits amonts du nœud qui reçoit l'association. D'autre part, les traits avals du nœud pied de l'arbre auxiliaire doivent s'unifier avec les traits avals du nœud recevant l'adjonction.

Voici deux exemples simplifiés de substitution et d'association avec unification des traits :



La même structure précédente avec un modifieur (adjectif : *double*) peut donner le résultat suivant :



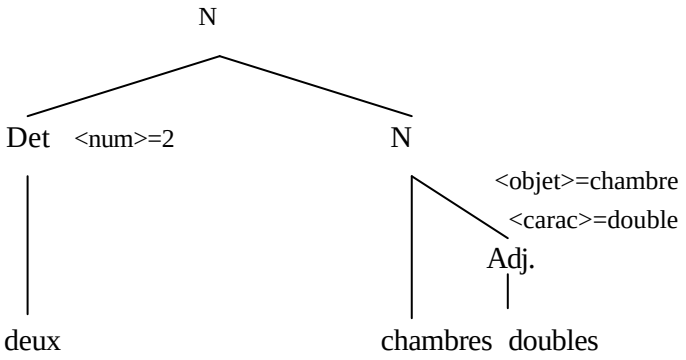


Figure 42. Exemples de substitution et d'association avec unification des traits

2.4.3.1.2 La propagation sémantique

La propagation sémantique est une opération qui porte sur les non-terminaux des arbres d'analyse. Elle vise principalement à mieux intégrer les catégories sémantiques et les catégories syntaxiques associées aux branches des différents arbres élémentaires. Au départ, les arbres locaux sont associés chacun à une catégorie sémantique simple qui correspond à son rôle dans le discours. Au fur et à mesure de l'analyse, la représentation sémantique associée aux arbres d'analyse s'enrichit. Cet enrichissement se fait selon deux mécanismes de base : la propagation prédicative et la propagation inductive.

- **Propagation prédicative :** elle consiste à monter la racine d'un arbre vers la racine de l'arbre qui le domine. Elle est utilisée notamment pour les connecteurs discursifs et pour les représentations sémantiques des éléments qui ne font pas partie d'une articulation thématique. Le schéma général de cette opération est présenté dans la figure suivante :

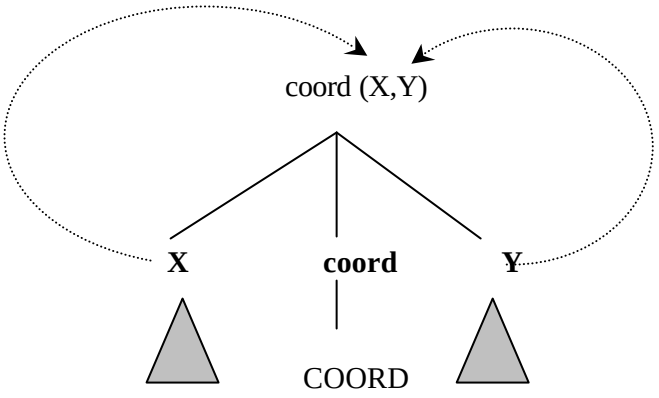
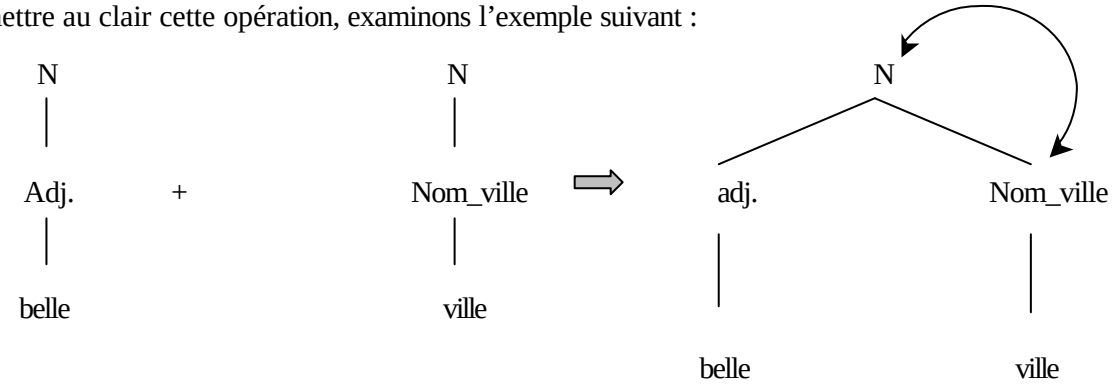


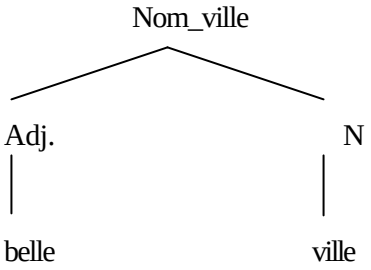
Figure 43. Schéma général de l'héritage simple

Comme nous pouvons le voir dans l'exemple précédent, les catégories racines des arbres ombrés sont propagées chacune vers sa place spécifique au sein d'un prédicat correspondant à la structure sémantique de la construction.

- **Propagation inductive :** les racines des arbres lexicaux initiaux sont converties de manière inductive en catégories sémantiques. Cela permet d'intégrer les arbres construits selon des critères syntaxiques (essentiellement ceux construits avec l'opération d'association) avec les arbres locaux et globaux qui sont basés principalement sur une catégorisation sémantique. Pour mettre au clair cette opération, examinons l'exemple suivant :



Après la propagation inductive de la catégorie intermédiaire de l'arbre qui a reçu l'association *Nom_ville* nous obtenons :



L'arbre ainsi obtenu peut être substitué à un nœud d'un arbre local comme s'il était un arbre lexical simple dont la racine est *nom_ville*.

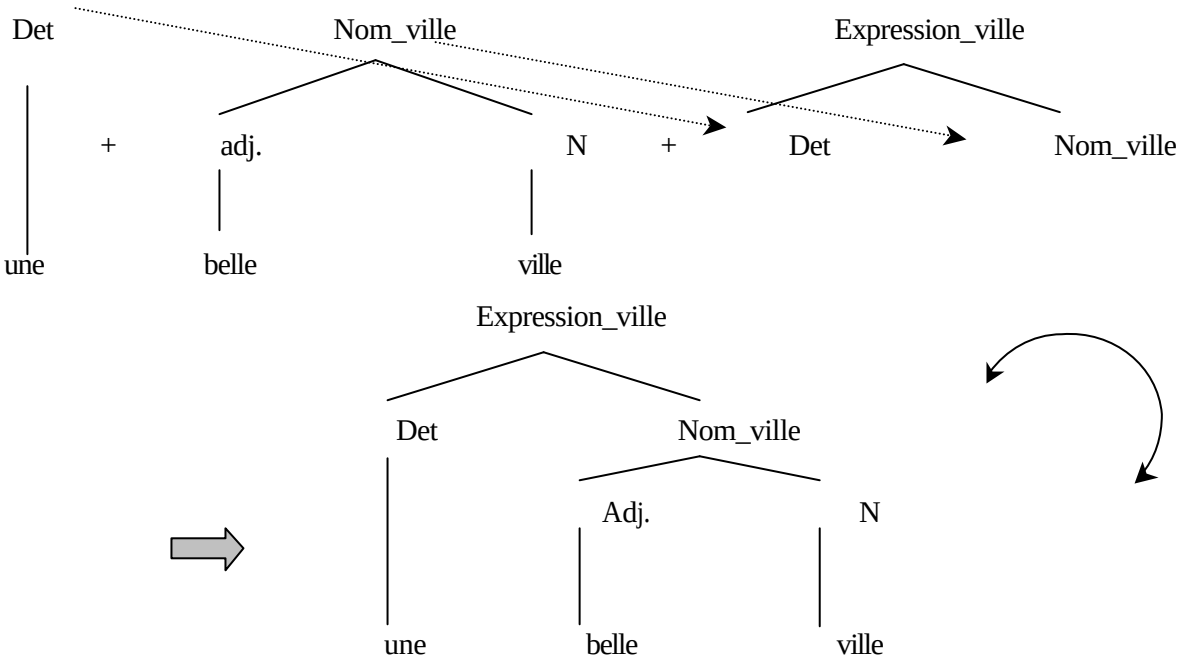


Figure 44. Exemple de propagation inductive

Ces deux mécanismes sont complètement indépendants de l'unification. En effet, contrairement à l'unification qui se passe simultanément aux opérations syntaxiques l'association et la substitution, la propagation sémantique est faite *a posteriori* sur le résultat obtenu avec les opérations syntaxiques ainsi que l'opération d'unification. Par ailleurs, il faut noter que l'opération de propagation prédicative est associée à l'opération de substitution alors que l'opération de propagation inductive est liée à l'opération d'association.

2.4.4 Exemples de traitement avec la Sm-TAG

Dans ces paragraphes nous allons nous concentrer sur deux problèmes précis :

1. Montrer la portée de la Sm-TAG en terme de traitement des phénomènes syntaxiques complexes.
2. Montrer l'adéquation des solutions proposées dans le cadre de la Sm-TAG avec les travaux dans le domaine de la syntaxe formelle notamment en ce qui conern l'effet de l'interaction directe entre la syntaxe et la sémantique sur le changement potentielle des relations de dépendance entre les constituants de l'énoncé.

Notre étude porte sur deux phénomènes linguistiques que nous avons jugés importants pour un formalisme de traitement du langage oral. Il s'agit de la négation et de l'emphase. Le choix de ces deux phénomènes est motivé par plusieurs raisons que nous allons discuter plus loin.

2.1.1.55 Méthodologie

L'objectif principal de notre étude étant de montrer la manière dont on traite les formes clés des phénomènes visés (la négation et l'emphase). Nous avons décidé de commencer d'abord par l'établissement d'une typologie de ces phénomènes[30]. Ainsi, afin de garantir à la fois la couverture des phénomènes complexes et les occurrences réelles de ces phénomènes dans les dialogues réels, nous avons jugé bon d'établir cette typologie sur la base de deux sources d'informations qui sont à la fois différentes et complémentaires :

1. Les typologies et les grammaires existantes : cette source nous a permis en particulier de couvrir les différentes constructions et formes des phénomènes visés tant sur le plan de l'écrit que sur celui de l'oral. Différents ouvrage de référence ont été utilisés pour cette tâche. Nous pouvons en citer : (Gadet, 1992), (Gadet, 1989), (Blasco-Dulbecco, 1999), (Blanche-Benveniste, 1990), (Blanche-Benveniste, 1997). Par ailleurs, nous avons utilisé plusieurs articles de recherche qui portent sur des points particuliers. Les références de ces articles seront citées au cours de la présentation.
2. Analyse de corpus : l'analyse de corpus de dialogue oraux nous a permis d'observer les occurrences orales des phénomènes visés et leurs divergences possibles avec les descriptions des phénomènes. Trois corpus ont été utilisés pour cette tâche : le corpus de réservation hôtelière du laboratoire CLIPS-IMAG (Hollard, 1997), le corpus Murol (Caelen *et al.*, 1997), le corpus du projet DALI (Sabah, 1997).

La typologie obtenue a été codée sous forme de règles syntaxiques dont le nombre total est de 137 règles : 32 règles pour l'emphase et 105 pour la négation. Ces règles ont servi de base pour générer (à la main) un corpus de 252 énoncés contenant les différentes structures décrites par la grammaire (voir (Kurdi et Ahafhaf, 2002) inclus à l'annexe de cette thèse, pour plus de détails sur le processus de génération).

Ainsi, dans notre étude nous avons pris ce corpus comme une base pour extraire **les cas clés** qui nous semblent intéressants de discuter en ce qui concerne la couverture et la profondeur de la Sm-TAG.

2.1.1.56 La négation

Nous allons commencer ce paragraphe par une discussion de l'intérêt de la négation pour notre étude. Nous allons ensuite passer à la présentation des différents éléments de la négation ainsi que leurs variations formelles et fonctionnelles.

2.4.4.1.1 Intérêt de la négation

L'intérêt de la négation par rapport à notre étude est pluridimensionnel. En effet, ce phénomène combine différentes propriétés intéressantes à la fois pour l'oral et la Sm-TAG :

---

[30] Cette typolgie a été réalisée en collaboration avec notre collègue Mohamed Ahafhaf (voir (Kurdi et Ahafhaf, 2002)).

- La négation est un phénomène qui est à la fois sémantique, syntaxique et lexical. En effet, d'un point de vue sémantique la négation est équivalente à un opérateur qui inverse la valeur de vérité d'une proposition. D'un point de vue syntaxique, la négation implique différentes structures grammaticales qui interfèrent parfois avec d'autres phénomènes comme la coordination ou l'ellipse. Sur le plan lexical, la négation, selon les cas implique l'utilisation de termes appartenant à différentes catégories morphologiques : adverbes, déterminants, pronoms, etc. Ainsi, c'est un phénomène particulièrement intéressant pour notre formalisme qui combine les niveaux sémantiques, syntaxique et lexical dans le même cadre.
- La négation est un phénomène qui présente des particularités intéressantes à l'oral. En effet, comme nous avons vu dans la première partie de cette thèse, la négation est considérée par certains chercheurs comme l'archétype de déviance de l'oral par rapport à la syntaxe de l'écrit.

2.4.4.1.2 Le terme ne

Le terme *ne* est l'un des deux éléments qui sont généralement utilisés pour marquer la négation : le terme *ne* couplé avec un autre élément dont la nature peut varier selon le type de la négation. Par ailleurs, ce terme peut être séparé du deuxième élément par d'autres mots ou constituants (un syntagme verbal généralement) comme dans :

(54) je **ne** voudrais **pas** une chambre simple

De plus, le terme *ne* précède directement le deuxième élément de la négation dans les constructions infinitives :

(55) Il m'a demandé de **ne pas** annuler la réservation

Finalement, le *ne* tout seul peut indiquer dans certains cas la négation. Il s'agit du *ne* dit littéraire qui utiliser avec certains verbes comme : cesser, pouvoir, oser, etc.

(56) Il ne cesse de parler

Comme son nom l'indique le *ne* littéraire est utilisé dans les œuvres littéraires. Par ailleurs, ce terme est aussi utilisé à l'oral.

Outre son emploi comme un élément de négation, le mot *ne* peut être utilisé dans diverses constructions tant dans des textes littéraires que dans les dialogues oraux. Dans ce cas, il est appelé le *ne* explétif. Trois types de contextes peuvent être distingués :

1. Le *ne* qui précède un certain nombre de verbes (comme craindre, douter empêcher nier, etc.).

   (57) Elle a **peur** qu'il **ne** revienne

2. Le *ne* avec des conjonctions : le schéma général de ces constructions est le suivant : conjonction pronom *ne* verbe :

(58) C'est possible **à moins que** la chambre **ne** soit réservée

En dehors de *à moins que*, différentes conjonctions peuvent précéder le terme *ne*. Nous pouvons en citer : *avant que*, *de peur que* et *de crainte que*.

3. Le *ne* couplé avec des comparatifs : le mot *ne* peut être couplé avec un comparatif comme dans l'exemple suivant :

(59) La chambre est plus chère que je **ne** le pensais

Sur le plan fonctionnel, comme nous avons dit dans les paragraphes précédents, la possibilité de suppression du mot *ne* à l'oral constitue l'un des principaux points de divergence entre la syntaxe de l'écrit et celle de l'oral. Cependant, l'élision du *ne* n'est pas toujours possible à l'oral et comme nous avons vu dans certains cas, le *ne* tout seul peut marquer la négation. Cette souplesse d'utilisation a rendu le statut du *ne* un sujet de débat au sein de la communauté de linguistique française. En effet, il existe trois possibilités pour analyser le terme *ne* :

- Le premier de ces courants (voir par exemple (Corblin, 1995) et (Abeillé et Godard, 1997)) considère le *ne* comme un clitic jouant le rôle de l'affixe du verbe et donc n'étant pas vraiment une partie de la négation. Outre la possibilité de son élision, ce groupe prend comme argument le fait que le terme *ne* est utilisé dans des constructions non négatives comme celles que nous venons de voir avec le *ne* explétif.
- Le deuxième courant stipule que le terme *ne* est un élément de la négation (voir (Muller, 1991)). Dans notre étude nous nous inscrivons dans le cadre de ce courant pour les deux raisons suivantes :
  i. Comme nous avons vu, dans certains contextes, le terme *ne* peut tout seul exprimer la négation. Cela veut dire que ce terme joue un rôle direct dans le marquage de la négation.
  ii. A l'oral, la suppression du *ne* n'est pas possible dans les contextes où ce terme joue un rôle de désambiguïsation syntaxique comme dans la négation et la coordination de deux syntagmes verbaux :

  (60) il **ne** mange **ni ne** boit rien

  Comme nous pouvons le constater, le premier *ne* dans l'énoncé précédent n'est pas facultatif que ça soit à l'oral ou à l'écrit puisqu'il sert à délimiter l'étendue de la négation. Ce point fera l'objet d'une discussion plus approfondie plus loin dans la section des conjonctions négatives.
- Outre ces deux possibilités précédentes, le mot *ne* peut être vu comme la première composante d'un morphème discontinu dont la deuxième partie est le deuxième élément de la négation (pas, point, etc.). Selon cette possibilité le morphème de la négation a la forme suivante :

  ne …. Second élément (pas, point, rien, etc.).
  ⌊_____⌋

Comme nous avons vu, cette analyse n'est pas compatible avec les données réelles. En effet, d'une part, nous avons observé un bon nombre de cas où l'un des deux éléments de la négation est facultatif ou parfois impossible. D'autre part, comme nous allons le voir plus loin, les mots qui peuvent jouer le rôle du second élément de la négation ont des formes et des fonctions syntaxiques assez variées (adverbes, pronoms, déterminants, etc.). Cela réduit la possibilité de l'existence d'un morphème unique.

Ainsi, sur le plan syntaxique, nous considérons le terme *ne* comme étant un élément à part entière dans l'énoncé et qui joue un rôle parfois central dans les constructions négatives. Par contre, sur le plan discursif, le terme *ne* ne joue pas un rôle particulier dans l'énoncé (il ne peut pas être thème ou rhème par exemple). Par ailleurs, ce terme n'a pas un effet direct sur le changement des rôles thématiques au sein de l'énoncé. Cela limite son traitement au niveau des arbres lexicaux au sein du formalisme Sm-TAG.

Ainsi, dans le contexte d'analyse par la Sm-TAG, nous sommes devant trois configurations qui nécessitent des techniques différentes de traitement :

1. **La configuration générique (ne verbe pas) :** dans cette configuration le mot *ne* joue le rôle d'adverbe de négation au même titre que le mot *pas* (ou n'importe quel mot qui peut être à sa place). Le traitement de cet adverbe est similaire à celui des autres adverbes. Voici comme exemple le traitement du segment (…) *ne réserve pas* :

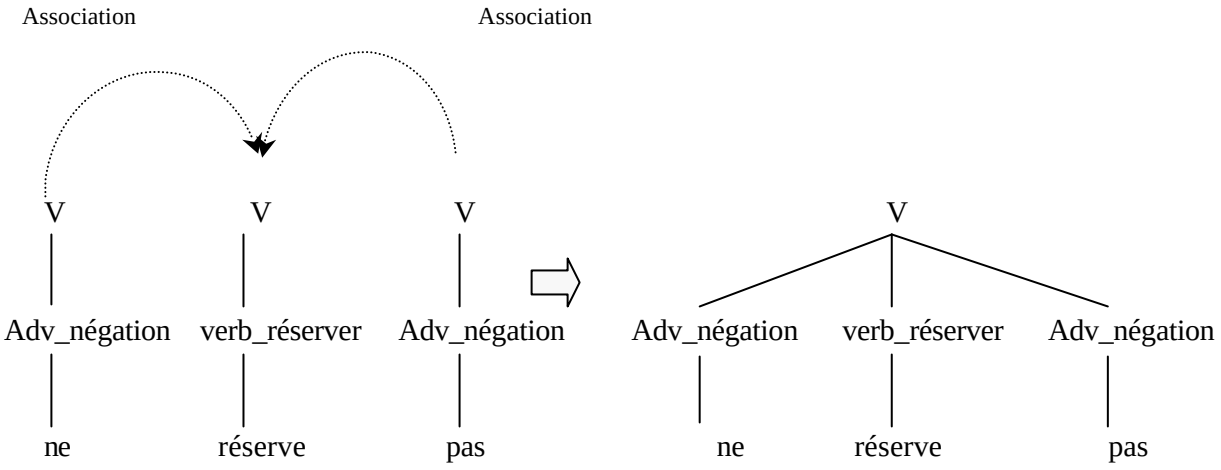


Figure 45. Exemple de traitement du ne dans la configuration générique

2. **La configuration infinitive :** dans ce cas, comme nous avons vu, le terme *ne* et le second élément de la négation se mettent devant le verbe formant une locution négative : pour montrer la manière dont les locutions de négations (ne pas) sont traitées dans le cadre de la Sm-TAG, prenons comme exemple le segment d'énoncé suivant et son traitement : **ne pas** réserver….

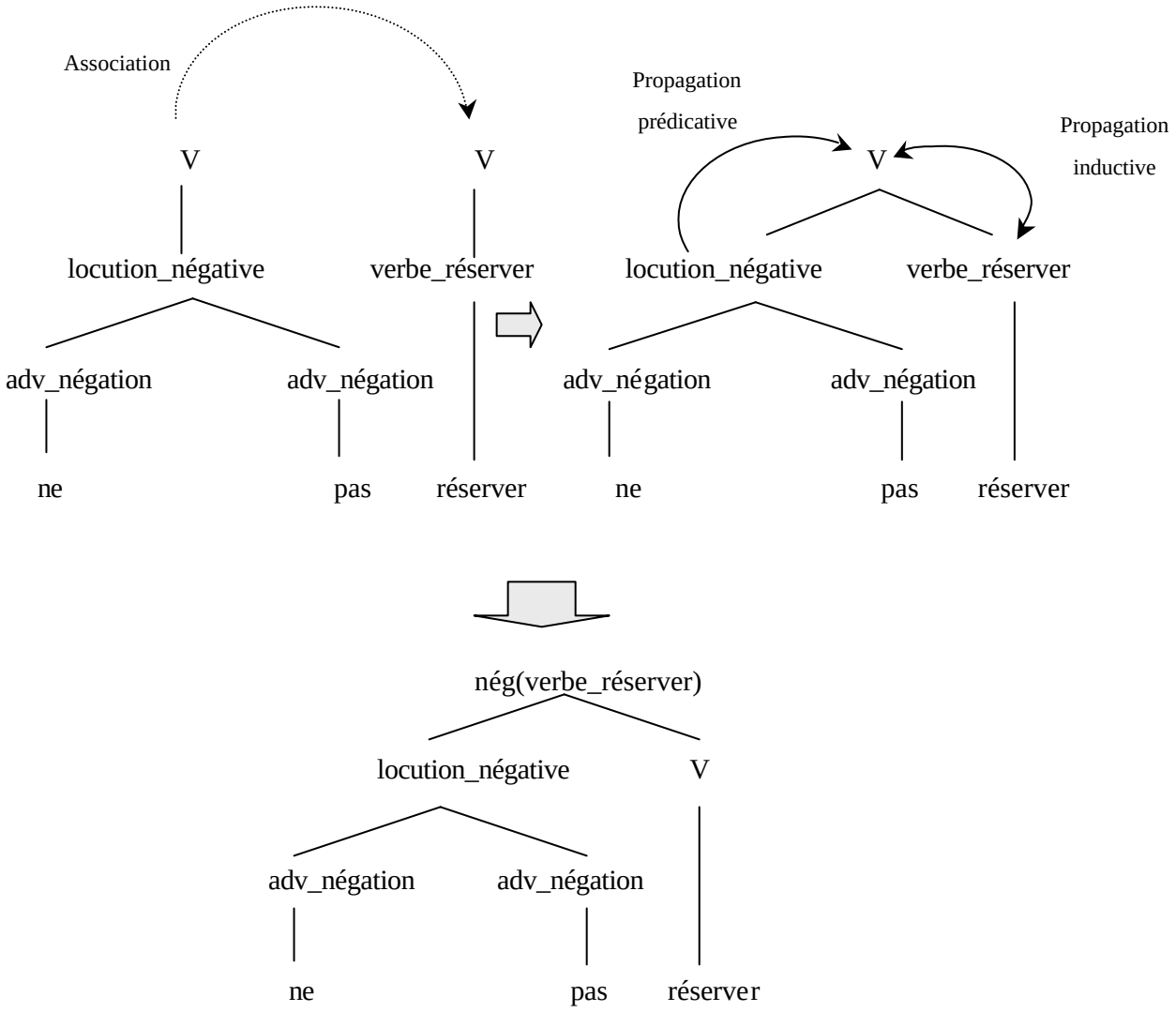


Figure 46. Exemple de la négation d'un verbe infinitif

Comme nous pouvons le voir dans l'exemple précédent, la locution négative qui est représentée par un arbre couvre les deux éléments de la négation (*ne* et *pas*). Par ailleurs nous avons vu que cet arbre a pour racine la catégorie verbe. Cela permet de le lier à n'importe quel prédicat verbal à l'aide de l'opération d'association.

3. Le ***ne*** seul : nous avons vu que le mot *ne* peut être utilisé tout seul dans deux cas : le *ne* littéraire et le *ne* explétif. Le traitement de ces deux termes n'est pas fondamentalement différent de celui des adverbes en général. En effet, chacun de ces deux termes est représenté par un arbre lexical dont la racine est la catégorie *verbe* qui lui permet de s'associer aux prédicats verbaux de l'énoncé. Pour résoudre l'ambiguïté possible entre ces deux termes qui ont des comportements

syntaxiques et sémantiques différents, les racines des arbres qui les représentent sont enrichies par des traits sémantiques indiquant la nature de l'adverbe que le verbe peut prendre pour marquer la négation. Cette distinction est possible étant donné que les groupes de verbes qui peuvent être modifiés par chacun de ces deux adverbes sont des groupes fermés et dont les membres peuvent être délimités facilement. Voici les arbres utilisés pour le *ne* explétif et le *ne* négatif :

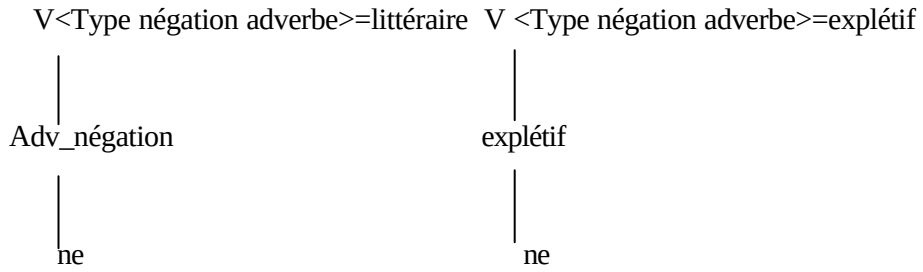


Figure 47. Les arbres lexicaux utilisés pour la représentation du *ne* explétif et du *ne* négatif

Après l'association, la catégorie *Adv_négation* est propagée à la racine de l'arbre crée contrairement à la catégorie *explétif*.

2.4.4.1.3 Les adverbes de négation

Selon nos observations des trois corpus que nous avons utilisés dans notre étude, la forme la plus fréquemment utilisée comme deuxième élément de négation est les adverbes de négation. Sur le plan syntaxique, il existe trois manières pour présenter les relations de ces adverbes avec le verbe en français (Abeillé et Godard, 1997) :

- La première consiste à utiliser des catégories fonctionnelles supérieures au verbe.
- La deuxième consiste à le traiter au même niveau que le verbe au sein du syntagme verbal.
- Finalement, la troisième consiste à adjoindre l'adverbe directement au verbe.

Nous estimons avec (Williams, 1994) que la troisième possibilité est la meilleure à la fois à cause de sa simplicité formelle et à cause des différentes données empiriques qui montrent que le comportement de l'adverbe de négation n'est pas fondamentalement différent de celui des autres adverbes en français. En effet, dans cette langue, comme le notent (Di Sciullo et Williams, 1987) les différents types d'adverbes peuvent s'adjoindre à droite du verbe. Voici, à titre d'exemple, l'arbre d'analyse syntaxique de l'énoncé : *je ne voudrais pas une chambre*.

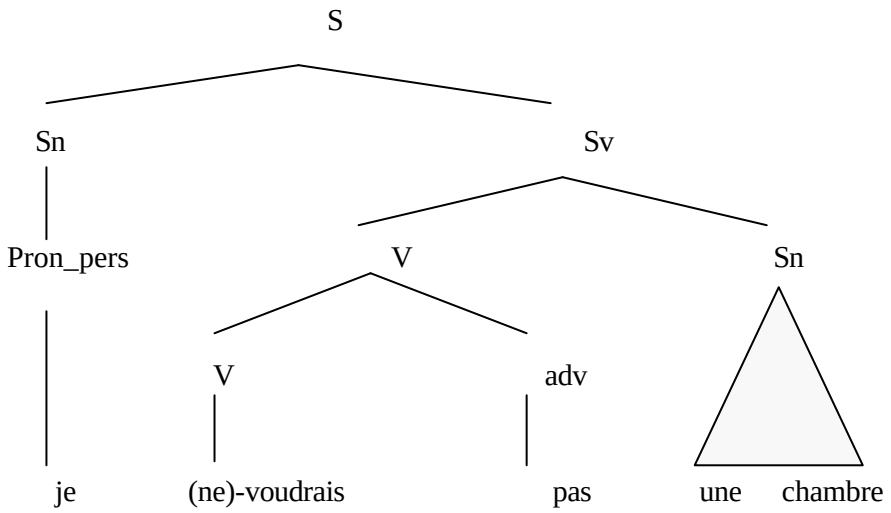


Figure 48. Arbre syntaxique correspondant à l'énoncé négatif

Comme nous pouvons le voir dans la figure précédente, l'adverbe est traité comme un modifieur du verbe et il est ainsi directement associé à lui.

Sur le plan discursif, les adverbes de négation ont un comportement similaire à celui du terme *ne*. En effet, les adverbes de négation ne jouent pas un rôle thématique particulier dans l'énoncé et n'ont pas un effet direct sur la distribution des rôles thématiques. Cependant, ces adverbes servent généralement à délimiter des segments qui jouent un rôle thématique. Voici un exemple de deux énoncés respectivement affirmatif et négatif analysés d'un point de vue discursif :

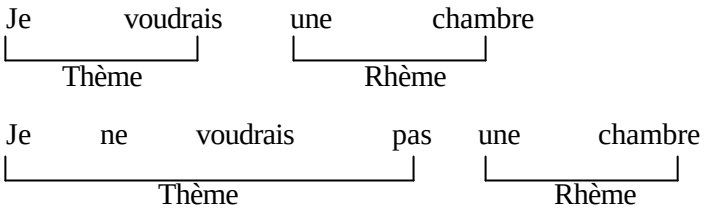


Figure 49. Analyse discursive d'un énoncé avec un adverbe de négation

En ce qui concerne le traitement dans la Sm-TAG, bien que la priorité dans ce formalisme soit donnée aux critères sémantiques plutôt qu'aux critères syntaxiques, il n'existe pas de conflit entre la syntaxe et la sémantique par rapport à la relation entre le verbe et l'adverbe. Pour mettre au clair cette idée, examinons le traitement de l'énoncé : *je ne voudrais pas une chambre*, dans le cadre de la Sm-TAG.

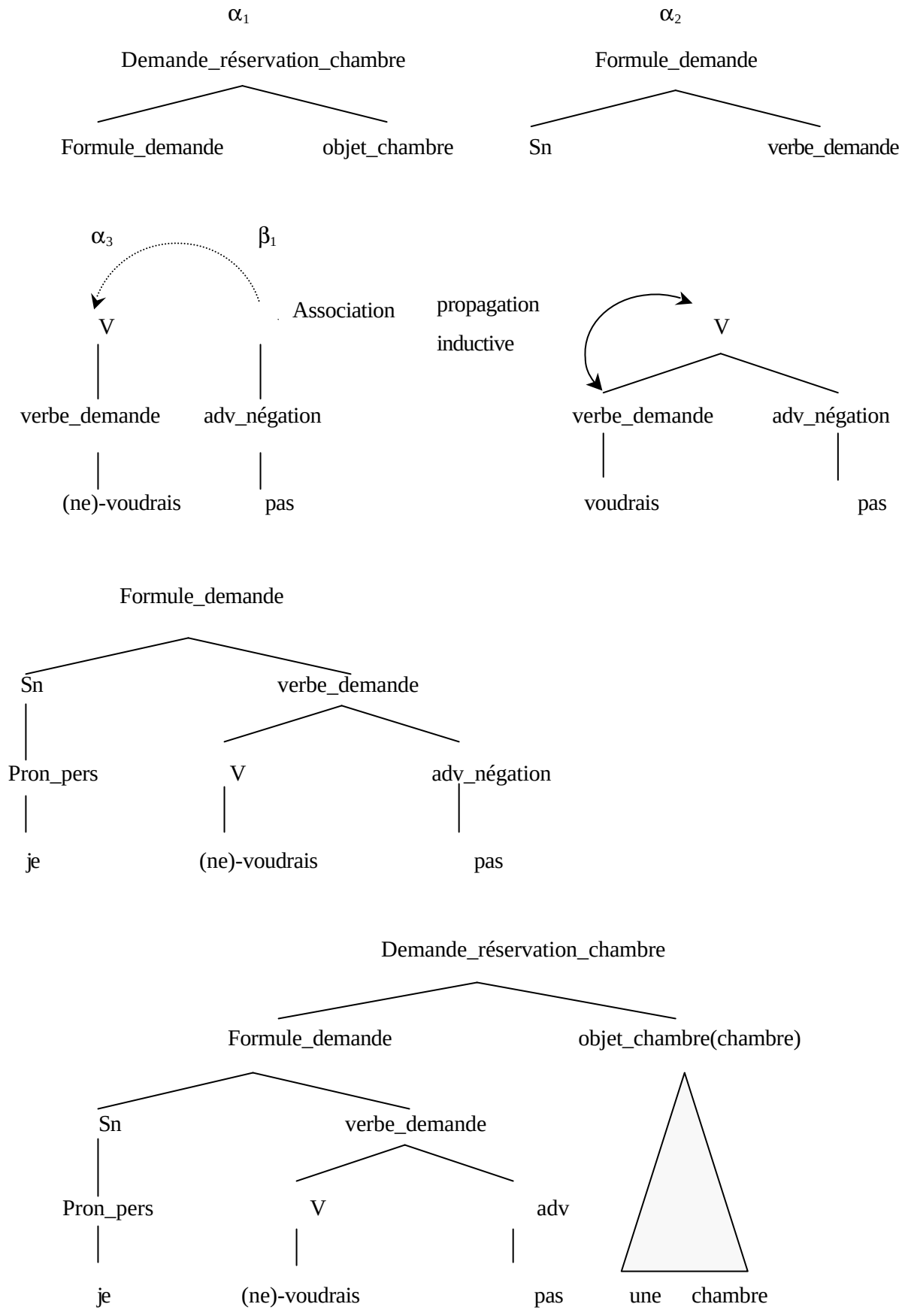


Figure 50. La représentation d'une construction négative dans la Sm-TAG

Comme nous pouvons le remarquer dans les deux arbres d'analyse, l'arbre d'analyse syntaxique et l'arbre d'analyse par la Sm-TAG, le traitement de l'adverbe de négation est fait pratiquement de la même manière dans les deux cas. En effet, dans l'arbre Sm-TAG le verbe et l'adverbe appartiennent au même niveau d'analyse et sont dominés directement par le même constituant supérieur : le constituant *verbe_demande* joue le même rôle que le constituant Sn dans l'arbre syntaxique.

2.4.4.1.4 Les déterminants de négation

Contrairement aux adverbes de négation, les déterminants de négation agissent sur les constituants nominaux. Ces éléments sont parfois appelés déterminants indéfinis (Riegel *et al.*, 1994), ou même adjectifs de négation. Sur le plan sémantique, les déterminants de négation indiquent qu'il n'existe pas d'occurrence dans l'univers référentiel pertinent qui vérifie le prédicat.

En français, il existe différents déterminants de négation tel que : aucun, nul(le), pas un(e), pas un(e) seul(e), etc. Comme nous pouvons le remarquer dans la liste précédente, nous pouvons distinguer entre deux types d'adjectifs : des adjectifs simples (aucun, nul) et des adjectifs composés (pas un, pas une seule, etc.).

1. **Les déterminants simples :** il s'agit de mots qui déterminent directement la tête du syntagme nominal comme nous pouvons le voir dans l'arbre d'analyse syntaxique du segment : *aucune chambre*.

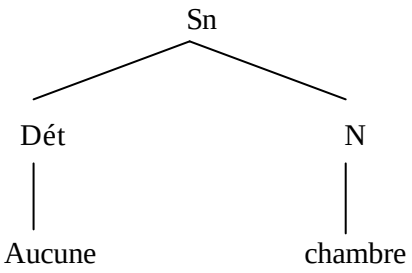


Figure 51. Arbre syntaxique partiel représentant la place d'un déterminant de négation au sein d'un syntagme nominal

Sur le plan discursif, tout comme les adverbes de négation, les déterminants de négation ne jouent pas un rôle thématique dans l'énoncé. Ainsi, dans le cadre de la Sm-TAG, ces éléments sont traités à l'aide d'arbres lexicaux seulement. Par ailleurs, le traitement de ces éléments est similaire à celui des déterminants en général : substitution du déterminant au nœud correspondant dans l'arbre local. Voici à titre d'exemple le traitement de l'énoncé : *je ne voudrais aucune chambre*.

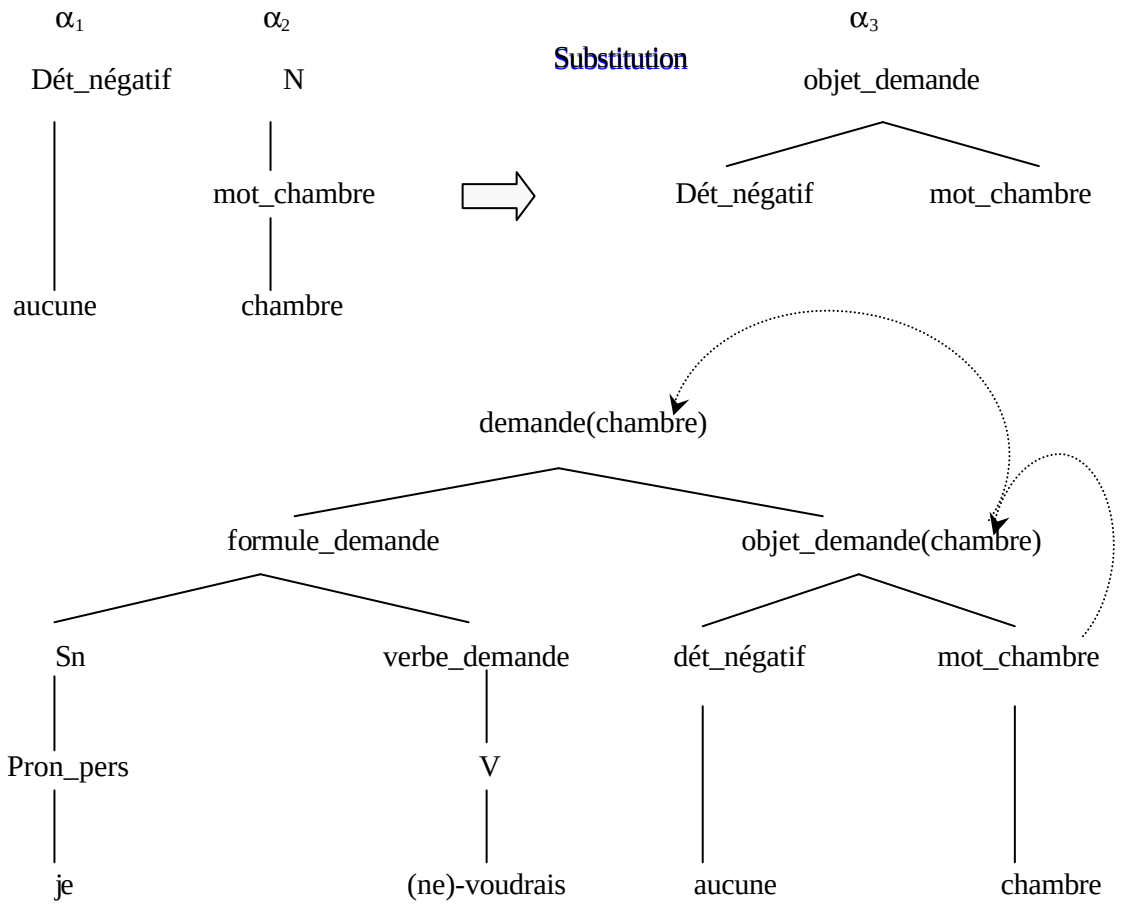


Figure 52. Le traitement des déterminants négatifs simples dans le cadre de la Sm-TAG

Comme nous pouvons le remarquer dans les arbres précédents, la seule différence entre l'arbre Sm-TAG et l'arbre d'analyse syntaxique réside dans la nature du constituant qui groupe le déterminant négatif et le nom : objet_demande dans l'arbre Sm-TAG vs. Sn dans l'arbre d'analyse syntaxique. Cela montre que les dépendances syntaxiques sont respectées dans ce phénomène.

2. **Les déterminants composés :** le problème principal lié au traitement de ces déterminants comparés aux précédents est leur aspect semi-figé. En effet, la négation est constituée d'une locution dont le deuxième élément est un déterminant qui, comme tout autre déterminant, s'accorde en genre et nombre avec le nom qu'il détermine.

(61) Il (ne) connaît **pas un seul** hôtel.

(62) **Pas une** chambre n'a été prise.

Pour résoudre le problème de ce semi-figement tout en conservant la cohérence de traitement, ces locutions sont traitées avec un arbre dont la tête est le déterminant :

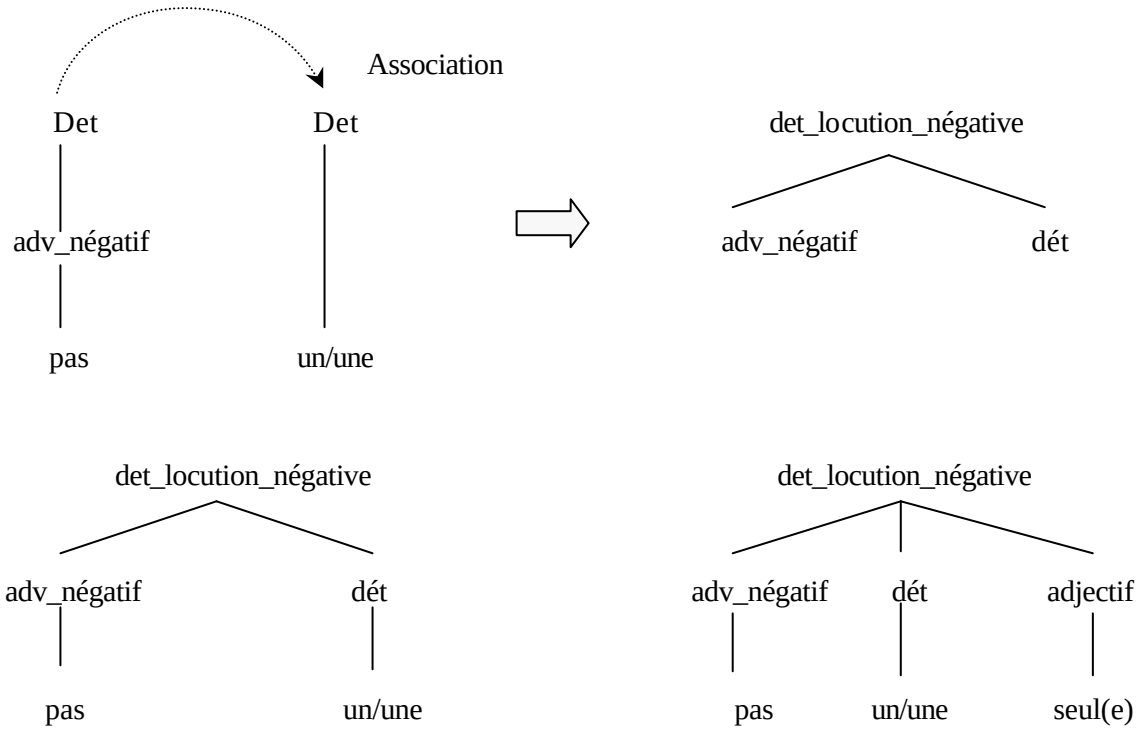


Figure 53. Les arbres utilisés pour le traitement des locutions négatives

Les arbres ainsi construits peuvent être associés à l'arbre d'analyse à l'aide de l'opération de substitution. Ainsi, cette solution permet de combiner les deux avantages suivants :

1. Sur le plan linguistique, elle permet de prendre en considération la particularité de la relation entre les mots formants la locution en les groupant au sein d'un même arbre.
2. D'un point de vue pratique, elle constitue une solution acceptable qui permet de prendre en considération la souplesse de ces constructions liée notamment à l'accord du déterminant avec le nom.

2.4.4.1.5 La conjonction négative

Lorsque la négation porte sur plus d'un syntagme ou groupe nominal, des conjonctions négatives sont utilisées. Le schéma général de ce type de négation est le suivant : ni constituant$_1$ ni constituant$_2$. Ainsi, deux grands types de constructions négatives avec des conjonctions peuvent être distingués : la coordination de constituants verbaux et la coordination de constituants non verbaux. Pour la simplicité de l'exposé, nous allons commencer par la présentation du deuxième type.

1. **Coordination des constituants non-verbaux :** comme nous pouvons le voir dans les exemples suivants, les conjonctions négatives peuvent coordonner différents types de constituansts non-verbaux :

(63) **Ni François ni Pierre** ne sont venus (sujets, Sn)

(64) Il **n**'est arrivé **ni tôt ni tard** .. (complément circonstanciel de temps)

(65) je **ne** voudrais **ni une chambre ni une suite** (objets, Sn)

Comme nous pouvons remarquer dans les exemples précédents, le terme *ne* est obligatoire avant les verbes. D'un point de vue discursif, chacun des éléments coordonnés joue un rôle thématique particulier dans l'énoncé. Prenons à titre d'exemple la structure discursive de l'énoncé 65 qui est présentée dans la figure suivante :

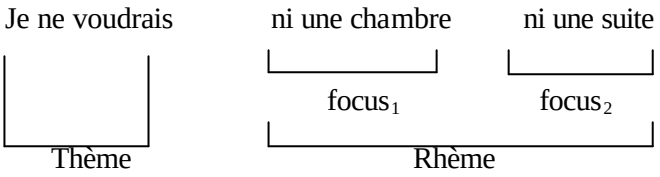


Figure 54. Analyse discursive d'un énoncé avec des conjonctions négatives

Comme nous pouvons le voir dans la figure précédente, les deux éléments coordonnés jouent chacun le rôle de focus et constituent ensemble le rhème de l'énoncé.

Pour discuter les aspects syntaxiques des conjonctions négatives examinons l'arbre d'analyse syntaxique de l'énoncé 65 :

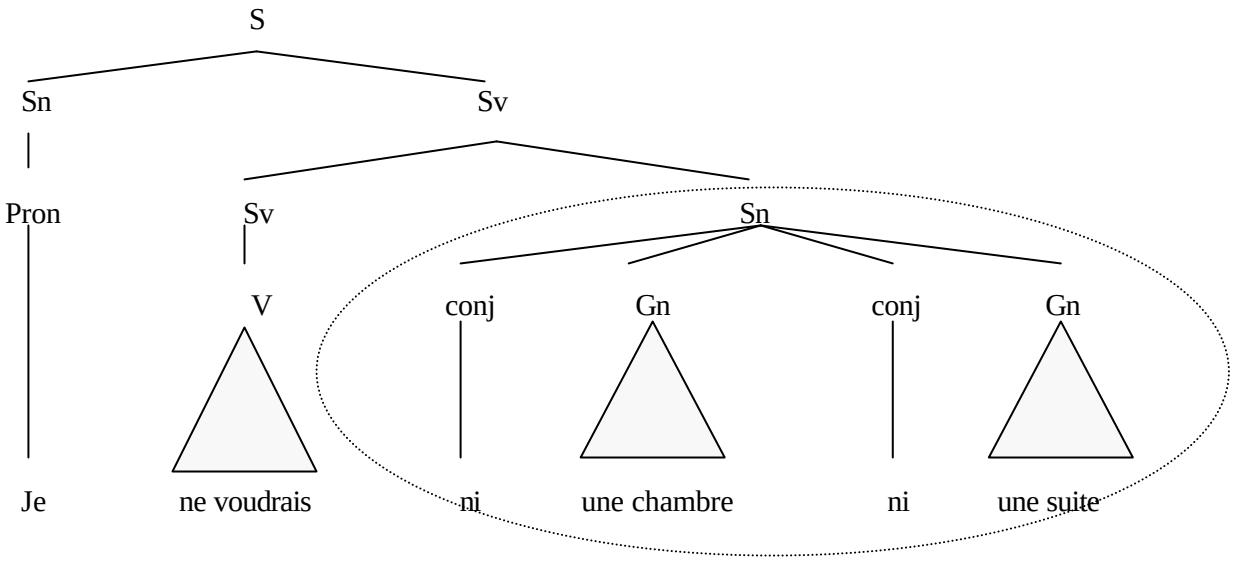


Figure 55. Arbre d'analyse syntaxique d'un énoncé avec des conjonctions négatives

Si nous comparons l'arbre d'analyse syntaxique à l'analyse discursive de l'énoncé 65, nous pouvons noter que les constituants coordonnés ont dans les deux cas un rôle identique. De même, dans les deux cas, les constituants coordonnés dépendent d'un même constituant de niveau supérieur (Rhème dans la structure discursive et Sn dans la structure syntaxique).

Dans le cadre de la Sm-TAG, un arbre spécial est utilisé pour traiter ce genre de constructions :

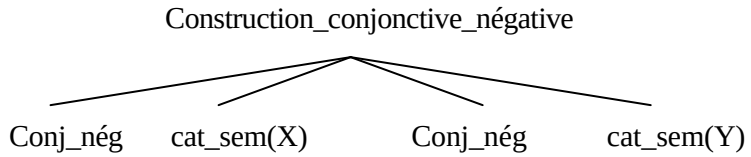


Figure 56. Arbre de base pour le traitement de la coordination des constituants non-verbaux

Deux remarques peuvent être faites à propos de cet arbre :

- Les dépendances syntaxiques des différentes composantes de la construction sont respectées même si les catégories des racines de ces constructions ne sont pas syntaxiques. En effet, nous remarquons que les conjonctions négatives dépendent directement de la construction principale de coordination (Construction_conjonctive_négative) au même titre que les arbres coordonnés.
- Cet arbre est valable quelle que soit la fonction syntaxique des éléments coordonnés (sujet, objet direct ou objet indirect) ou leur structure (participe, adjectif, Sn, Sp, etc.).

Pour concrétiser ces idées prenons comme exemple l'arbre d'analyse suivant :

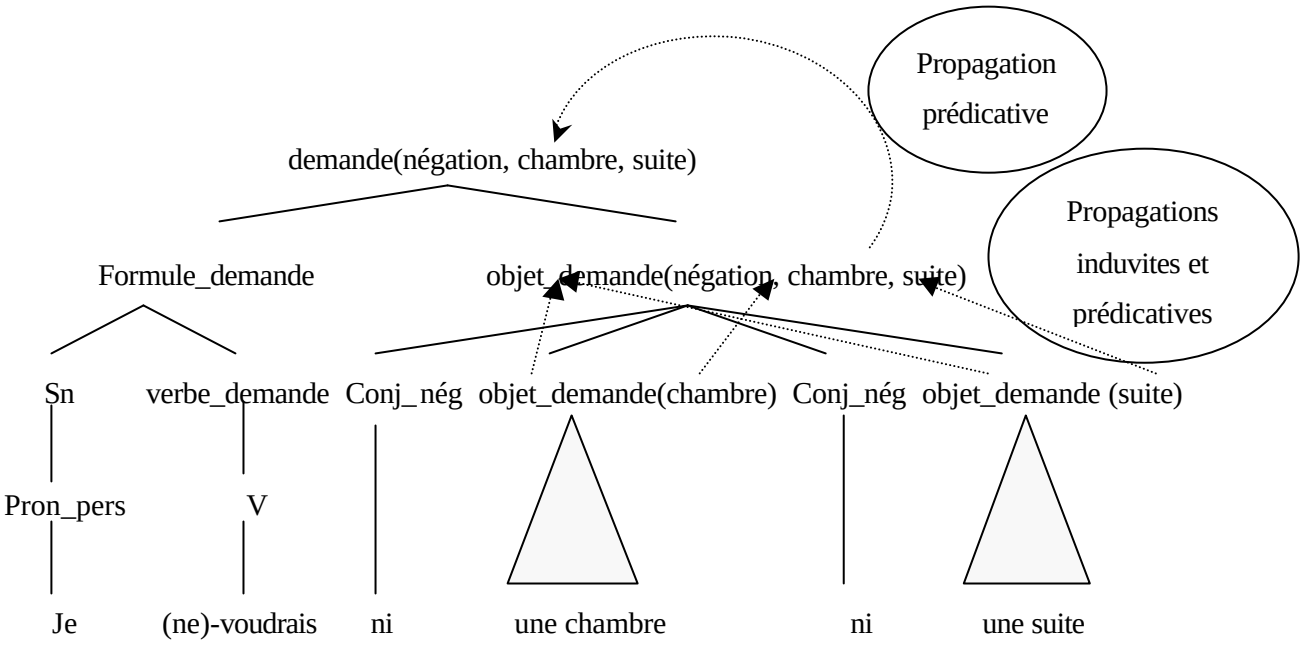


Figure 57. Arbre d'analyse d'un énoncé avec des conjonctions négatives

Comme nous pouvons le remarquer dans l'arbre précédent, la catégorie racine des deux arbres coordonnés *objet_demande* a été propagée à la racine de l'arbre de la construction de coordination (les deux éléments coordonnés ont toujours la même catégorie racine dans les cas que nous avons observés dans notre corpus). Ceci est fait à l'aide de l'opération de propagation inductive. Par ailleurs, les arguments des racines des arbres coordonnés sont propagés à l'aide de

l'opération de propagation prédicative à la racine de la construction de coordination et ensuite ce même contenu est propagé aussi à l'aide de l'opération de propagation prédicative à la racine de l'arbre d'analyse. La racine finale obtenue correspond à la structure sémantique globale de l'énoncé analysé.

2. **La coordination de constituants verbaux :** ce type de construction consiste à coordonner deux constituants verbaux avec une conjonction négative.

(66) Il **ne** **parle ni** **ne** **lit le russe**

La différence principale avec la coordination des constituants non-verbaux, comme nous pouvons le constater dans l'exemple précédent, est que nous avons une seule conjonction de coordination plutôt que deux. Par ailleurs, comme nous avons deux constituants verbaux, le terme *ne* se répète deux fois : une fois devant chaque constituant verbal. Pour tester la possibilité de supprimer l'un de ces deux termes, nous avons demandé à des locuteurs natifs de juger la grammaticalité des trois énoncés suivants :

(67) Il parle ni lit le russe*

(68) Il parle ni **ne** lit le russe*

(69) Il **ne** parle ni lit le russe ≈ (à la limite de l'acceptable)

Comme nous pouvons le voir dans les énoncés précédents, les deux cas où le premier *ne* est supprimé ont été jugé agrammaticaux par les sujets. Ce jugement est motivé par l'ambiguïté créée par l'absence du premier *ne*. En effet, dans ce cas les sujets s'attendent à un complément du premier prédicat verbal qu'ils jugent affirmatif mais à sa place ils trouvent une conjonction négative et un autre prédicat verbal. En ce qui concerne le troisième cas, il est jugé à la limite de l'acceptable puisque la présence du terme *ne* avant le premier prédicat verbal permet de savoir qu'il s'agit d'une construction négative. L'absence de symétrie entre les deux constituants verbaux est le point qui rend cette possibilité à la limite de la grammaticalité. En effet, notre observation de symétrie syntaxique et discursive dans la coordination des constituants non-verbaux est valable ici aussi.

D'un point de vue traitement avec la Sm-TAG, le traitement de ces constructions se fait avec l'arbre suivant :

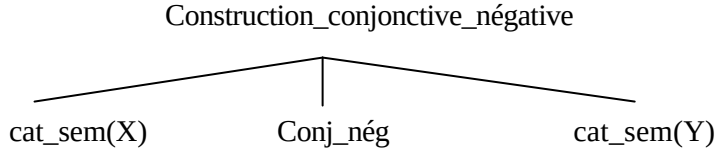


Figure 58. Arbre de base pour le traitement de la coordination des constituants verbaux

Le mécanisme de traitement de la coordination des constituants verbaux est assez similaire à celui des constituants non-verbaux. En effet, dans ce cas aussi nous utilisons les opérations de propagation inductives et prédicatives pour enrichir progressivement la construction obtenue.

2.1.1.57 L'emphase

2.4.4.1.6 Intérêt de l'emphase

La mise en emphase est un procédé qui consiste à attribuer une importance particulière à une partie de l'énoncé émis par le locuteur. Le choix de ce phénomène est motivé par les raisons suivantes :

1. C'est un phénomène grammatical qui a une implication forte sur les niveaux sémantiques et discursifs de l'énoncé. Cela lui donne un intérêt particulier pour notre formalisme qui intègre ces différents niveaux.
2. La mise en emphase se fait selon différents mécanismes liés à l'ordre des mots dans l'énoncé qui joue un rôle particulièrement important dans le cadre de la Sm-TAG pour l'attribution des fonctions syntaxiques et des rôles sémantiques et discursifs.

En français, deux moyens syntaxiques sont possibles pour mettre un élément en emphase[31] : la dislocation et l'extraction.

2.4.4.1.7 La dislocation

Ce moyen consiste à détacher un constituant en tête ou en fin de l'énoncé avec une reprise avec un pronom. En effet, la dislocation est associée à un double marquage où l'élément détaché est remplacé par une apposition (un pronom) qui contribue à la mise en focus de l'élément remplacé. Prenons les deux énoncés suivants :

(70) Je prends **la chambre**

(71) **La chambre** je **<u>la</u>** prends

Comme nous pouvons le remarquer dans les deux énoncés précédents : le syntagme nominal *la chambre* détaché au début de l'énoncé a été aussi marqué par le pronom *la*.

La dislocation prend différentes formes selon la nature de l'élément détaché ou celle de l'apposition utilisée. Ainsi, nous pouvons distinguer entre deux principaux types de dislocation :

4. **Détachement d'un syntagme nominal :** le syntagme nominal peut être détaché en tête ou en fin de l'énoncé. Les fonctions des syntagmes détachés sont assez variés comme nous pouvons le voir dans les exemples suivants :

---

[31] Outres ces deux moyens syntaxiques, le français dispose de l'accent d'insistance pour mettre un élément en emphase. L'accent d'insistence peut mettre en valeur des éléments linguistiques de types variés : sujet, verbe, la tête du syntagme nominal objet, etc. Ce moyen n'a pas été retenu dans notre étude étant donné qu'il ne met pas en œuvre des transformations syntaxiques ou sémantiques qui permettent de montrer un trait particulier de la Sm-TAG.

(72) **Ces chambres <u>elles</u>** sont bonnes (Sujet)

(73) **La réservation cela/ça/c**'est important (sujet)

(74) **Ces chambres** je **<u>les</u>** prends (complément d'objet direct)

(75) **Ce séjour** ma femme **<u>en</u>** rêve (complément d'objet indirect)

(76) **Dans cet hôtel** on (**<u>y</u>**) trouve des belles chambres (complément circonstanciel de lieu)

Comme nous pouvons le voir dans les énoncés précédents, l'élément détaché est repris par un pronom clitique ou démonstratif pour garder l'ordre canonique des éléments de l'énoncé. La seule exception à cette règle est le détachement du complément circonstanciel de lieu qui, selon la grammaire normative, ne doit pas être repris par un pronom. A l'oral, l'obligation ou l'interdiction de reprise par un pronom ne sont pas toujours respectées (Riegel, 1994). En effet, les compléments obligatoires à l'écrit sont parfois omis à l'oral et vice-vera, le complément circonstanciel de lieu est parfois repris par le pronom *y*.

Sur le plan discursif, cette dislocation consiste à inverser les rôles thématiques des constituants. En effet, lorsqu'un thème est détaché, il devient un rhème ou vice-versa.

Syntaxiquement, le détachement consiste à modifier la dépendance du syntagme nominal par rapport au verbe de l'énoncé. Prenons l'exemple suivant pour discuter concrètement cette idée :

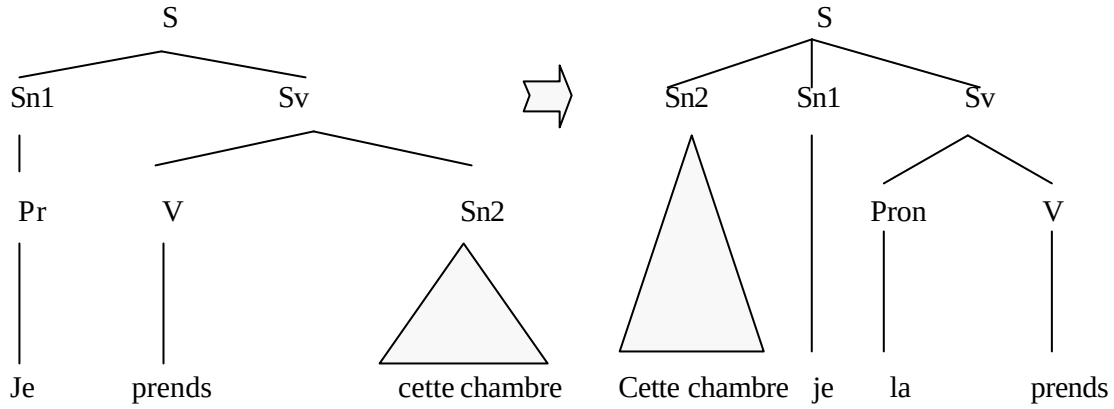


Figure 59. Analyse syntaxique d'un énoncé avec une dislocation d'un syntagme nominal objet

Comme nous pouvons le remarquer dans la figure précédente, le rattachement du Sn2 est différent dans l'énoncé sans dislocation de celle de l'énoncé avec dislocation. En effet, ce syntagme étant une apposition du pronom objet *la* on pourrait penser qu'il doit dépendre du syntagme verbal au même titre que le pronom objet *la*. Une telle analyse viole la règle de continuité des éléments de l'arbre d'analyse étant donné que le pronom sujet *Sn1* sépare *Sn2* et *Sv*. Par ailleurs, sur le plan sémantique, l'attachement du *Sn1* à la racine de l'arbre directement permet d'exprimer sa distinction par rapport au reste de l'énoncé.

Ainsi, nous avons adopté l'analyse présentée dans la figure précédente où le syntagme nominal détaché dépend directement de la racine de l'arbre d'analyse. L'analyse obtenue dans le cadre de la Sm-TAG est présentée dans la figure suivante :

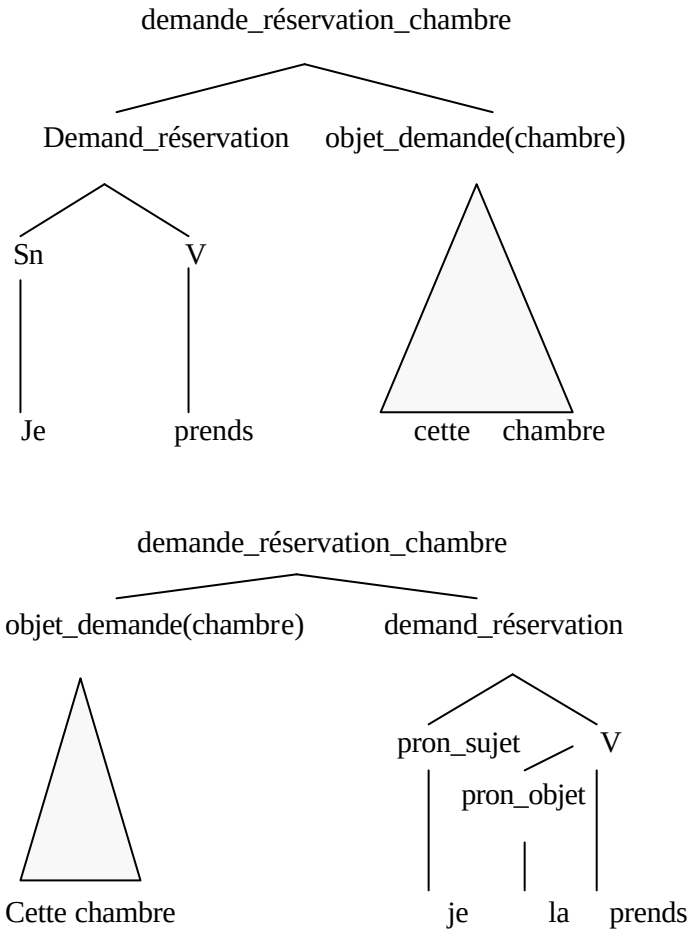


Figure 60. Analyse d'un énoncé avec extraction dans le cadre de la Sm-TAG

Nous pouvons remarquer deux points dans les deux arbres d'analyse précédents comparés aux arbres syntaxiques des mêmes énoncés :

i. La Sm-TAG permet de garder cette spécifié du thème par rapport au reste de l'énoncé en permettant de l'attacher directement à la racine de l'énoncé.

ii. Le traitement du pronom personnel, attaché au verbe à l'aide de l'opération d'association, dans l'arbre d'analyse Sm-TAG est assez similaire à celui de l'arbre d'analyse syntaxique du même énoncé.

2. **Détachement de groupes infinitifs et de propositions subordonnées complétives :** le détachement des groupes infinitifs et des complétives se fait pratiquement dans les mêmes conditions. En effet, dans les deux cas, il est associé aux pronoms personnels ou démonstratifs, dont la répartition dépend de la même fonction syntaxique. Le détachement des groupes

infinitifs qui jouent la fonction *sujet* est la forme principale de détachement des groupes infinitifs ou des complétives. L'usage des groupes infinitifs et des complétives comme sujet n'est pas très fréquent même si cela est possible d'un point de vue grammatical. Ainsi, dans l'usage courant, le détachement de ces constructions est la forme préférée (Riegel, 1994). Dans ce cas, seuls les pronoms démonstratifs ou impersonnels peuvent être utilisés pour la reprise du groupe détaché. Par ailleurs, tout comme les groupes nominaux, les groupes infinitifs et les complétives détachés peuvent être en début ou en fin de l'énoncé. Voici quelques exemples :

(77) réserver maintenant **Cela m'ennuie**. (groupe infinitif/début)

(78) que Frank ait réservé **Cela amuse Françoise**. (complétive/début)

(79) **C'est dommage** que la chambre soit réservée. (complétive/fin)

(80) **Il est dommage** que la chambre soit réservée. (complétive/fin/pronom impersonnel)

Sur le plan discursif, ce type de détachement consiste en l'inversement des emplacements des éléments occupants des rôles thématiques. En ce qui concerne la Sm-TAG, le traitement de ces phénomènes est assez similaire à celui des cas de détachement des syntagmes nominaux. En effet, les pronoms démonstratifs ou impersonnels sont traités de la même manière que les pronoms clitiques ou démonstratifs utilisés pour reprendre les syntagmes nominaux.

2.4.4.1.8 L'extraction

Ce phénomène consiste à associer un présentatif et un relatif pour extraire un constituant de la phrase et qui permet d'obtenir les clivées. Par ailleurs, une construction similaire dite semi-clivée peut être associée aux phénomènes d'extraction. En effet, cette construction combine l'extraction et le détachement d'un constituant pour le mettre en emphase.

1. **Les clivées :** le clivage est l'un des principaux moyens d'emphase en français. Il consiste en l'emploi des présentatifs *c'est….. qui / que* qui encadrent, en le plaçant en tête de phrase, l'élément mis en emphase qui peut être de natures diverses. En effet, comme nous pouvons le voir dans les exemples suivants, l'extraction peut affecter des sujets (clivée sur l'agent), des objets (clivée sur le patient), des compléments circonstanciels, etc. (Riegel, 1994).

   (81) C'est le client qui réserve la chambre. (clivée sur l'agent)

   (82) C'est la chambre que réserve le client. (clivée sur le patient)

   (83) C'est demain que j'arrive (complément circonstanciel)

   Sur le plan discursif, la mise en emphase peut porter sur le thème (clivées sur le l'agent) ou sur le rhème (clivées sur le patient ou sur le complément circonstanciel).

L'élément verbal des présentatifs peut varier en temps et en mode. Cette variation reste cependant facultative.

(84) C'est avec une C.B. que j'ai payé.

(85) C'est avec une C.B. que je paye.

L'accord entre d'une part le groupe nominal mis en emphase par les présentatifs et le verbe est facultatif à l'oral. En effet, comme nous avons vu dans l'introduction de ce chapitre, le non-respect de l'accord dans les clivées constitue le cas prototypique du non respect de l'accord à l'oral en français.

(86) Ce **sont** les clients qui réservent.

(87) C'**est** les client**s** qui réservent.

Par ailleurs, l'extraction est un phénomène qui s'inscrit dans une problématique plus vaste qui est celle de l'ordre des mots (Blasco-Dulbecco, 1999). En effet, outre le déplacement du groupe mis en relief à la tête de la phrase, le pronom *que* donne la liberté de changement de l'ordre entre le syntagme nominal sujet et le verbe de la phrase : sujet verbe vs. verbe sujet. Par contre, cette variation n'est pas possible avec *qui*.

(88) C'est la chambre que réserve le client.

(89) C'est la chambre que le client réserve.

(90) C'est le client qui réserve la chambre.

(91) C'est le garçon qui la chambre réserve. (agrammatical)

D'un point de vue traitement dans le cadre de la Sm-TAG, trois points peuvent être notés :

i. L'accord n'étant pas une source d'information retenue dans le cadre du formalisme Sm-TAG, les cas d'extraction où l'accord n'est pas respecté ne posent pas un problème particulier pour le traitement avec ce formalisme.

ii. Les extractions portent sur des éléments qui jouent un rôle discursif particulier (en général thème ou rhème). Ainsi, l'adoption des unités discursives (plutôt que syntaxiques) comme base de l'analyse dans la Sm-TAG nous permet de capter toutes les subtilités des extractions.

iii. Les présentatifs (c'est … qui/que) sont la seule partie qui nécessite un traitement particulier au sein de la Sm-TAG. Comme nous l'avons vu, il s'agit d'éléments auxiliaires dont la fonction est la mise en emphase du groupe encadré. La structure résultante (*c'est* élément_encadré *que/qui*) a les mêmes propriétés syntaxiques et

sémantiques que celles de l'élément encadré. Ainsi, nous avons jugé bon de traiter ces présentatifs comme une construction semi-figée qui hérite sa représentation sémantique de l'élément encadré.

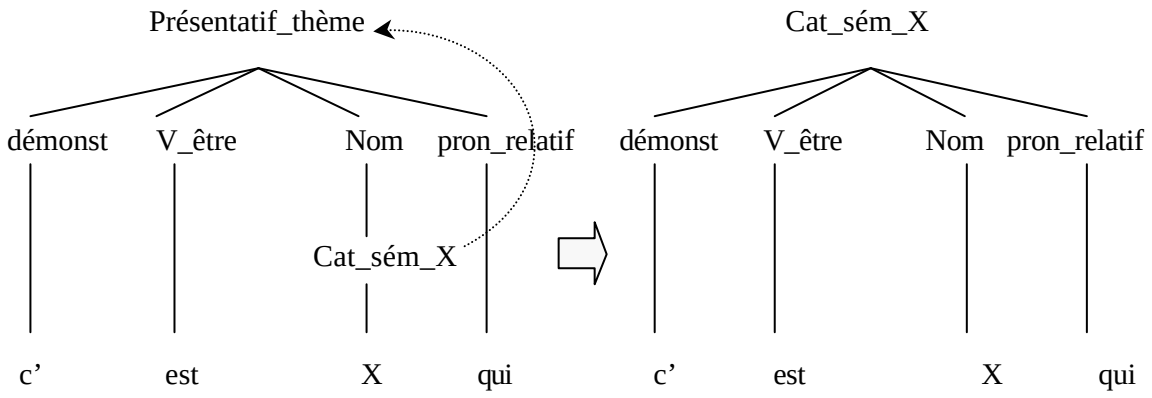


Figure 61. L'arbre de base pour le traitement des présentatifs et le processus d'héritage sémantique

Comme nous pouvons le remarquer dans la figure précédente, l'arbre utilisé est indépendant de la structure encadrée qui peut être un syntagme nominal, syntagme prépositionnel, complément circonstanciel, etc. Cela permet d'utiliser ces arbres pour traiter les présentatifs quelle que soit l'application dans laquelle ces arbres sont utilisés.

2. **Les pseudo-clivées :** les énoncés pseudo-clivés sont séparés en deux parties ; introduite par *ce que*, la première partie consiste généralement en une relative périphrastique alorsque la deuxième partie qui est introduite par *c'est* peut consister en un groupe nominal, infinintif ou une complétive.

(92) **Ce que** je voudrais **c'est** une chambre. (groupe nominal)

(93) **Ce que** je désire **c'est** de réserver une bonne chambre. (infinitif)

(94) **Ce que** je veux **c'est** que vous me trouviez une chambre. (complétive)

Le rôle des présentatifs utilisés dans les semi-clivées se limite à la distinction des éléments mis en emphase du reste de l'énoncé. Pour vérifier cette idée, il suffit de supprimer les présentatifs *ce que … c'es*t pour voir que l'énoncé obtenu est équivalant sémantiquement à l'énoncé avec les présentatifs (à l'exception de l'emphase elle-même bien entendu).

Par ailleurs, sur le plan discursif, les semi-clivées sont assez similaires aux clivées dans la mesure où l'élément mis en emphase est toujours un élément qui joue un rôle thématique particulier dans l'énoncé. Par contre, nous n'avons pas observé des cas d'inversement de position comme dans les clivées sur le patient. Le schéma général des semi-clivées est le suivant : ce que **Thème** c'est **Rhème.** Ainsi, les présentatifs peuvent être considérés, sur le plan

syntaxique, comme des expressions figées qui jouent le rôle de l'auxiliaire à la construction présentée par ces éléments.

Voici l'analyse proposée dans le cadre de la Sm-TAG pour le traitement des présentatifs utilisés dans les semi-clivées :

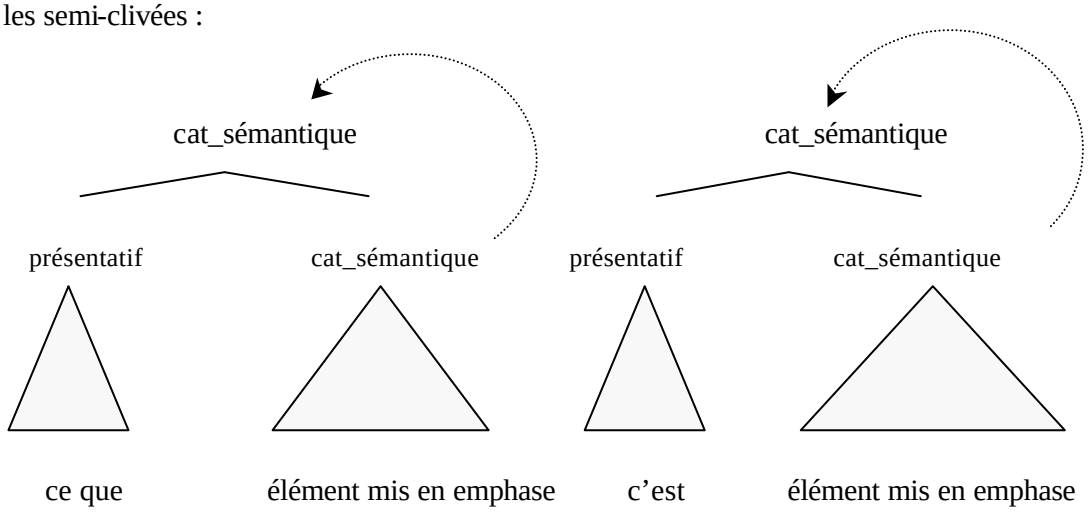


Figure 62. Le traitement des présentatifs des semi-figées dans le cadre de la Sm-TAG

Comme nous pouvons le remarquer dans la figure précédente, les présentatifs sont considérés comme des éléments auxillaires par rapport aux éléments présentés dont la nature sémantique délimite le contenu des éléments présentés. Par ailleurs, nous pouvons aussi remarquer que ces arbres ne dépendent pas de l'élément préssenté et peuvent ainsi être utilisés dans différentes grammaires destinées à des applications différentes.

2.4.5 La Sm-TAG : un formalisme pour l'analyse du langage oral

2.1.1.58 La Sm-TAG et l'architecture logicielle des modules d'analyse linguistique du langage oral

L'interaction des différentes sources de connaissances au sein du système d'analyse linguistique est l'un des problèmes centraux dans la conception et la réalisation de l'architecture de ce système. En effet, les interdépendances de ces différentes connaissances obligent parfois à utiliser des architectures complexes afin de prendre ces interdépendances en considération lors de l'analyse (Delmonte et Bianchi, 2002), (Mahesh, 1995). L'un des problèmes majeurs qu'affrontent les systèmes modulaires est la conversion (ou traduction) des informations émanant d'un module *A* en un format *compréhensible* par un module *B* qui a besoin de cette information pour prendre ses décisions. L'intégration des différentes sources de connaissances au sein du même cadre formel (comme la Sm-TAG) nous semble la meilleure solution de ce problème.

2.1.1.59 La Sm-TAG : un formalisme pour l'analyse robuste

La source principale du manque de robustesse d'un point de vue d'un formalisme linguistique est la complexité des traits à vérifier. En effet, plus un formalisme est complexe plus nous avons une chance d'avoir un problème lié à une contrainte dont la vérification par l'algorithme d'analyse est nécessitée

par ce formalisme. L'exemple le plus clair de ce genre de cas est le phénomène de l'accord. En effet, la vérification de l'accord du sujet avec le verbe nécessite le recours à des dispositifs spéciaux comme les traits d'accords ou autres. Le problème est que le système de reconnaissance peut créer des erreurs de reconnaissance partielles où le mot est reproduit avec la mauvaise marque d'accord. Ce cas est particulièrement valable pour le français où l'accord est marqué souvent par des *e* muets comme : arrivé**e**, réservé**e**, clair**e**, etc. Cela s'applique aussi aux verbes où c'est parfois impossible de distinguer phonétiquement les formes plurielles ou singulières d'un verbe comme arrives, arrive, arrivent, manges, mange, mangent, etc. Le principe de pertinence que nous avons adopté au sein de la Sm-TAG nous permet d'augmenter la robustesse de l'analyse en nous passons des sources de connaissances dont le rôle n'est pas central dans le traitement et qui peuvent au contraire constituer une source de bruit.

2.4.6 Discussion de la validité cognitive de la Sm-TAG

La faculté de langage fait partie intégrante du système cognitif général chez l'homme. Ainsi, tout modèle du langage humain doit être compatible avec les connaissances que nous avons sur la cognition humaine.

2.1.1.60 Un peu de méthodologie

Différents travaux dans le domaine de la psycholinguistique computationnelle présentent des approches algorithmiques ou linguistiques comme étant plausibles cognitivement (voir (Milward et Sturt, 1995), (Kaiser, 1999), comme exemples de ces travaux). Or, la notion de plausibilité cognitive reste à nos yeux assez vague et ne permet pas de donner un jugement précis sur l'adéquation des approches proposées avec la réalité cognitive. Ainsi, avant de commencer la discussion de la plausibilité cognitive de la Sm-TAG, nous allons commencer par la présentation de ce dont nous entendons. En effet, notre conception sur la validité cognitive peut être résumée dans les points suivants :

1. Nous savons formellement avec le théorème d'incomplétude de (Gödel, 1931) qu'un modèle parfait n'existe pas. Ainsi, sur le plan de la modélisation cognitive du traitement du langage, cela veut dire qu'il n'existe pas un modèle parfait du processeur linguistique humain. Autrement dit, les modèles dits pertinents cognitivement ne le sont que partiellement.
2. Nous savons à travers les différents travaux dans le domaine de la psychologie expérimentale et de la neurophysiologie (voir (Rosenbaum, 1987) pour une revue générale de ces travaux) que l'esprit-cervau humain traite l'information selon des techniques extrêmement variées et qui changent selon les besoins de la situation.
3. Comme il est impossible d'une part de proposer un modèle parfait du processeur humain et d'autre part vue la richesse des mécanismes de traitements utilisés par ce processeur il est presque difficile de proposer une approche qui ne soit pas partiellement pertinente cognitivement.

Ainsi, une approche plausible cognitivement est une approche dont **les traits clés** sont pertinents dans **le cadre contextuel** pour lequel cette approche est proposée.

Dans notre cas, discuter la plausibilité cognitive de la Sm-TAG revient à discuter l'adéquation de son trait principal (qui est le mode d'interaction entre la syntaxe et les connaissances de niveau supérieur) avec les résultats des travaux dans le domaine de la psycholinguistique expérimentale.

2.1.1.61 Discussion de la plausibilité cognitive de l'interaction directe de la syntaxe avec les connaissances de niveau supérieur

Comme nous avons vu dans les paragraphes précédents, le mode d'interaction directe entre la syntaxe et la sémantique est l'une des propriétés les plus caractéristiques de la Sm-TAG par rapport à la majorité des formalismes grammaticaux qui distinguent nettement entre le niveau syntaxique et les niveaux d'analyses de rang supérieur comme la sémantique et le discours. Un bon nombre de travaux dans le domaine de la psycholinguistique expérimentale a montré l'existence d'une stratégie d'intégration précoce des informations syntaxiques avec les informations de haut-niveau au cours du processus de compréhension[32].

Parmi ces travaux nous pouvons citer ceux de (Tyler et Marslen-Wilson, 1977) qui ont procédé à une expérience visant à vérifier si la sémantique intervient avant la fin de l'énoncé ou si au contraire elle intervient au cours du traitement en parallèle avec l'analyse syntaxique. Pour ce faire, ils ont utilisé des paires ambiguës adjectif-verbe tel que *Landing planes* dans des énoncés comme 95 :

(95) a. If you walk too near the runway, *landing planes*…[33]

b. I you've been trained as a pilot, *landing planes*….

Les résultats de cette expérience ont montré que la sémantique intervient avant la fin de l'énoncé. En effet, lorsque le mot *planes* était suivi par un mot approprié par rapport au contexte (comme *are* pour (95a)), le temps de réponse était moins long que dans les cas où il y avait dans le même endroit un mot inapproprié (comme *is* pour (95a)). Cela montre que les sujets ont une préférence sémantique (induite de leurs connaissances générales sur le monde ainsi que de l'analyse sémantique du début de l'énoncé)

---

[32] En fait, le mode d'interaction de la syntaxe et des connaissances de hat niveau est un sujet de controverse entre les spécialistes de la psycholinguistique expérimentale. Comme nous estimons avec ((Crooker, 1996) voir page 28) que la raison principale de ces controverses est la limitation des moyens actuels d'investigation scientifique (outils de détection des mouvements oculaires, outils d'imagerie cérébrale, limitations liées au contrôle des variables expérimentales, etc.) et pas les aspects inhérents à l'interaction entre la syntaxe et les connaissances de haut-niveau proprement dits (qui sont l'objet de notre travail), avons préféré d'éviter d'entrer dans ces débats et de nous limiter aux arguments en faveur de l'interaction directe de la syntaxe et de la sémantique.

[33] Nous avons jugé bon de donner les exemples des matériels linguistiques utilisés dans les différents travaux que nous présentons tels qu'ils sont (en anglais) afin de conserver toutes les propriétés linguistiques de ces matériels sans biais.

qu'ils appliquent pour choisir l'analyse syntaxique la plus plausible. Par ailleurs, (Crain et Steedman, 1985) ont utilisé des énoncés passifs avec des propositions relatives pour montrer que la sémantique ainsi que le contexte référentiel peuvent guider le choix de la structure syntaxique. Par exemple, avec des énoncés comme :

(96) a. The **teachers** taught by the Berliz method passed the test.

b. The **children** taught by the Berliz method passed the test.

Ainsi, les énoncés similaires à (96b) ont été jugés comme étant grammaticaux plus fréquemment que les énoncés du type (96a). La différence sémantique entre les deux énoncés semble être la raison de cet écart dans le jugement étant donné qu'il est plus probable qu'un enfant soit enseigné qu'un professeur. En outre, ces chercheurs ont montré que ces indices sémantiques interviennent avant la fin de l'énoncé ou même avant une frontière syntagmatique. Ainsi, (Crain et Steedman, 1985) concluent en considérant l'impasse d'analyse[34] comme un phénomène contextuel qui peut être évité par la connaissance du contexte dans lequel un énoncé est réalisé. Par exemple, selon ces chercheurs, les énoncés du type (96a) étaient jugés agrammaticaux parce que leur biais sémantique les éloignait de l'hypothèse d'une structure relative et les menait à une impasse d'analyse lorsqu'ils rencontrent le verbe *passed*. Cette impasse d'analyse a été évitée dans les énoncés similaires à (96b) où le contexte sémantique permet de guider les sujets vers une structure relative qui est la bonne syntaxiquement. Une étude similaire à celle de Crain et Steedman a été menée par (Trueszell et Tanenhaus, 1994) avec une perspective d'analyse du discours. En effet, les chercheurs ont utilisé des énoncés avec des verbes ambigus (dont la forme est identique aux participes passés équivalents) comme :

(97) a. The fossile examined….

b. The archeologist examined….

Comme nous pouvons le remarquer dans les énoncés précédents, le verbe examined a la même forme que le participe passé du même verbe. Ainsi, ces chercheurs ont trouvé comme (Crain et Steedman, 1985) que les connaissances sémantiques influencent directement le choix de la structure syntaxique et conduisent parfois à des impasses d'analyse.

(Carpenter et Just, 1988) ont procédé à une étude des mouvements oculaires de la lecture. Leurs travaux ont montré que la durée de fixation des mots anormaux sémantiquement par rapport au contexte était plus longue que celle de mots équivalents (en terme de longueur, fréquence, et

[34] L'impasse d'analyse est la traduction que nous proposons de l'expression *garden path*. Il s'agit des cas d'ambiguïté locale qui guident le processeur humain vers une analyse unique à partir de laquelle il est difficile ou parfois impossible de faire une correction de l'analyse (voir (Croocker, 1996) page 7, pour plus de détails sur ce phénomène).

adaptation syntaxique par rapport au reste de l'énoncé) mais normaux sémantiquement. Cela montre que l'analyse sémantique se fait en parallèle avec le processus de lecture (et donc avec l'analyse syntaxique).

Plusieurs expériences visant à clarifier le rôle du contexte discursif dans la compréhension ont été menées (Spivey-Knowlton et Tanenhaus, 1994), (Boland *et al.*, 1995). Dans une étude récente (Altmann, 1999) décrit deux expériences sur ce problème. Dans ces deux expériences les sujets devaient lire des énoncés du type :

(98) He drank some….

Ces énoncés ont été utilisés dans des contextes qui introduisent ou pas des objets potables. L'idée est qu'après le verbe *drank* les sujets sont supposés penser que l'énoncé n'a pas un sens si l'objet de ce verbe n'est pas un élément potable. Ainsi, après avoir demandé aux sujets d'examiner différents groupes d'énoncés, il a observé que les réponses négatives (c'est-à-dire que l'énoncé n'as pas de sens) nécessitent plus de temps lorsque le contexte antérieur. L'auteur conclut que les rôles sémantiques (agent, patient, récipient, etc.) associés aux arguments discursivement antérieurs d'un verbe (les arguments situés dans un tour de parole antécédent) sont sélectionnés au point de la tête verbale par les sujets en prenant en considération les rôles disponibles (qui n'ont pas été encore associés à un item lexical) même lorsque l'entité qui réfère explicitement à ces antécédents (les pronoms anaphoriques) est postverbale et que cette entité n'est pas encore traitée par les sujets. Cela montre, d'une part que le contexte discursif intervient dans l'analyse syntaxique d'un énoncé et que d'autre part, cette intervention se fait en parallèle avec l'analyse syntaxique étant donné que son effet est détecté avant la fin du traitement de l'énoncé.

2.1.1.62 Discussion de la validité de ces arguments par rapport à la Sm-TAG

Un bon nombre de chercheurs qui travaillent dans le domaine de la psycholinguistique soutient l'hypothèse de l'intégration immédiate des différentes sources de connaissances impliquées dans la compréhension : syntaxe, sémantique, connaissances sur le monde. Nous avons vu aussi que cette hypothèse a été validée à la fois sur des énoncés isolés que sur des énoncés ancrés dans un contexte discursif particulier. Cela nous permet de confirmer la validité de l'idée de l'intégration que nous avons adoptée dans le cadre de la Sm-TAG. Cependant, comme nous avons vu, les résultats expérimentaux ne nous permettent pas de savoir précisément la (ou les) stratégie(s) utilisée(s) par le processus humain pour combiner ces différentes sources de connaissance. Par conséquent, le mode d'intégration proposé par la Sm-TAG (tout comme les autres approches qui intègrent différentes sources de connaissances dans la compréhension comme celle de (McClealand et Kawamoto, 1986)) doit être vu comme une *métaphore* dont les bases sont plausibles cognitivement mais pas comme un modèle formel de ce mode d'intégration.

3 Conclusion de la deuxième partie

Dans cette partie, nous avons présenté nos deux études théoriques menées dans le cadre de cette thèse. Il s'agit du modèle des extragrammaticalités de l'oral que nous avons proposé sur la base de notre analyse du *Trains Corpus* ainsi que la formalisation de la grammaire sémantique et la proposition du formalisme Sm-TAG comme un cadre pour le traitement à la fois robuste et profond de l'oral.

En conclusion, nous allons établir un bilan général de ces deux études par rapport à l'état de l'art que nous avons présenté dans les deux premières parties de cette thèse :

3.1 Bilan de l'analyse des extragrammaticalités

Notre étude vise à modéliser les aspects syntaxiques des extragrammaticalités. Notre travail, à ce propos, se distingue par la proposition de schémas différents pour les différents types d'extragrammaticalités que nous avons observés dans notre corpus :

- Les extragrammaticalités lexicales : nous avons distingué, au sein de cette catégorie, plusieurs types comme les amalgames et les mots oraux.
- Les répétitions et les autocorrections : sur ce plan, nous avons proposé un modèle à base de patron, inspiré des travaux précédents (Shriberg, 1994), (Heeman, 1997). Par ailleurs, nous avons adopté une méthode d'étiquetage à deux niveaux qui permet de prendre en considération le contexte dans lequel un patron apparaît ainsi que les relations et les conflits éventuels qu'il peut y avoir entre les patrons.
- Les faux départs et les incomplétudes : nous avons proposé un schéma général qui segmente ces extragrammaticalités en un ensemble de zones qui jouent chacune un rôle particulier dans la détection et la délimitation. Cette distinction des différentes zones permet de contextualiser les segments mal formés et de réduire, par conséquent, le nombre des cas de surgénérations.

Notre étude contient une analyse détaillée des occurrences multiples et imbriquées des différentes formes d'extragrammaticalités au sein du même énoncé. Par ailleurs, nous avons pris en considération les fausses extragrammaticalités dans notre analyse du corpus afin de mettre l'accent sur l'aspect **sémantique** des extragrammaticalités.

3.2 Bilan de la S-TSG

La S-TSG est une formalisation que nous avons proposée de la grammaire sémantique classique. Les points clés qui distinguent la S-TSG de la grammaire sémantique classique sont les suivants :

- **Points théoriques :** ayant un statut linguistique et mathématique bien défini, la S-TSG est facilement comparables à d'autres formalismes et approches pour le traitement de l'oral.
- **Points pratiques :** la structuration de la S-TSG, selon trois niveaux d'unités : arbres lexicaux, arbres locaux et arbres globaux. Cela rend l'écriture et la modification de la grammaire une tâche plus facile comparée à celle avec la grammaire sémantique classique.

3.3 Bilan de la Sm-TAG

La Sm-TAG est un formalisme hybride qui intègre différents niveaux de représentation au sein du même cadre. Ceci est essentiellement dû à l'interaction directe de la syntaxe et de la sémantique au sein de ce formalisme qui est notamment réalisée grâce aux opérations d'association et de propagation sémantique. Les propriétés clés de ce formalisme peuvent être résumées dans les points suivants :

1. Contrairement à la grammaire sémantique classique, les différentes propriétés formelles et linguistiques sont analysées et bien connues.
2. Equivalence faible avec une CFG : cela facilite considérablement la tâche de l'analyse avec des algorithmes efficaces et rend la réalisation d'une version stochastique de ce formalisme une tâche réaliste.
3. Un modèle sémantique compact basé sur l'intégration de la notion de la pertinence dans la définition des traits. Par ailleurs, nous avons proposé deux opérations sur les non-terminaux de la grammaire qui facilitent l'intégration des arbres syntaxiques intermédiaires au sein des arbres sémantiques.
4. Adoption des unités discursives comme base de traitement : comme nous avons vu dans les exemples de traitement des phénomènes linguistiques avec la Sm-TAG, cela n'a pas constitué une limite pour traiter les différentes formes des phénomènes que nous avons abordés. En effet, nous avons vu que des phénomènes comme l'emphase ou la négation affectent uniquement les segments qui jouent un rôle thématique dans l'énoncé : on ne peut pas mettre en emphase ou nier un élément que nous jugeons comme marginal. Par ailleurs, nous avons vu qu'avec la Sm-TAG, nous pouvons traiter des cas syntaxiquement et sémantiquement complexes comme des énoncés avec à la fois des constructions négatives et des coordinations, et ce de manière simple.
5. Non-violation des relations de dépendance syntaxique : bien que la priorité principale dans la Sm-TAG soit donnée à la sémantique, les relations de dépendance dans les phénomènes syntaxiques (comme la négation) ont été conservées.
6. Généralisation : la généralisation est une propriété importante d'un formalisme linguistique, en particulier, pour un formalisme comme le nôtre qui intègre des connaissances dépendantes de la tâches (qui sont par définition non-généralisables). Nous avons vu dans les paragraphes précédents que les procédures utilisées dans la Sm-TAG pour traiter les différentes formes des phénomènes linguistiques considérés dans nos exemples sont indépendantes de la tâche. Cela

contribue à augmenter le pouvoir expressif de la grammaire (dans la mesure où nous pouvons couvrir plus de phénomènes avec un nombre relativement limité d'arbres et de règles d'inférences) ainsi qu'à augmenter sa portabilité (puisque nous disposons d'un noyau indépendant de la tâche qui peut être utilisé dans différentes applications).

7. Adéquation avec les résultats des travaux dans le domaine de la psycholinguistique expérimentale : l'interaction directe entre la syntaxe et la sémantique (qui est le trait distinctif principal de la Sm-TAG) est compatible avec les résultats de plusieurs travaux dans le domaine de la psycholinguistique expérimentale qui stipulent que la syntaxe et la sémantique interviennent en même temps au cours du traitement de l'énoncé.

Partie III : les systèmes Corrector, Safir, Oasis et Navigator pour l'analyse du langage oral

0 Introduction de la troisième partie

Après avoir dressé un bilan général des contraintes d'un système d'analyse linguistique du langage oral dans les deux premières parties de cette thèse et après avoir proposé des modèles pour les phénomènes grammaticaux et extragrammaticaux de l'oral, dans la troisième partie, nous allons nous consacrer dans cette partie à la réalisation de ces modèles afin de tester leur validité applicative.

Dans cette partie, nous allons présenter trois systèmes dans le cadre de deux axes applicatifs :

- **Traitement des extragrammaticalités :** sur cet axe, nous allons présenter le système Corrector. Il s'agit de l'implantation du modèle théorique des extragrammaticalités que nous avons proposé dans le premier chapitre de la troisième partie de cette thèse.
- **Analyse linguistique du langage oral :** nous allons, sur cet axe, présenter les systèmes SAFIR et OASIS qui sont respectivement des implantations que nous avons réalisées de la S-TSG et de la Sm-TAG. Le système SAFIR est un prototype destiné à faire une évaluation préliminaire de nos choix théoriques et pratiques alors que le système OASIS est conçu dans une optique d'intégration dans le cadre d'une application réelle qui est dans notre cas la traduction automatique de la parole.

1 Chapitre III.1 : Le système Corrector pour le traitement des extragrammaticalités du langage oral

Rappelons qu'il existe deux tendances diamétralement opposées dans la littérature. Selon la première, il est possible d'utiliser des techniques très superficielles à base de N-grams et de patrons pour traiter **tous** les phénomènes d'extragrammaticalités. Par ailleurs, les chercheurs qui suivent la deuxième tendance soutiennent que la syntaxe est absolument nécessaire pour le traitement et généralisent, par conséquent, son utilisation à tous les phénomènes. Or, comme nous avons dit dans notre discussion des différentes méthodes, il nous semble que certains phénomènes comme les extragrammaticalités lexicales, les répétitions et les autocorrections peuvent être traités avec des approches à base de patrons de manière plus simple et plus efficace qu'avec la grammaire, puisque ces phénomènes, par leur nature même, ne nécessitent pas d'informations syntaxiques profondes. Par contre, nous avons montré à l'aide d'exemples qu'avec des approches superficielles à base de N-grams, il est impossible de prendre en considération suffisamment de contexte pour traiter certains cas.

Par ailleurs, dans notre analyse du *Trains Corpus*, nous avons vu qu'il est possible de procéder à une modélisation syntaxique fine des faux-départs et des incomplétudes. Nous avons vu aussi que la prise en considération des dépendances entre les syntagmes constitue un facteur clé pour la détection de certains phénomène.

Ainsi, la solution idéale, à nos yeux, consiste à combiner les approches à base de patrons à celles d'analyse syntaxique afin d'optimiser le rapport coût de traitement/efficacité dans le traitement. Les informations sémantiques peuvent être aussi ajoutées à condition de ne pas rendre le système dépendant de la tâche.

Ainsi, nous présentons le système Corrector qui est basé sur l'intégration de techniques de reconnaissance de patrons, d'analyse syntaxique et sémantique superficielle.

1.1 Requis du système

Corrector est destiné à traiter les extragrammaticalités du langage oral c'est-à-dire à détecter la présence de ces phénomènes et à délimiter leur étendue dans l'énoncé.

Les principaux requis de notre système peuvent être résumés dans les points suivants :

- **Portabilité :** le système doit être utilisable non seulement dans différents domaines d'application (négociation de transport, réservation touristique, etc.) mais aussi il doit être facile à intégrer au sein de systèmes divers dont les composantes sont très différentes. Ainsi, Corrector

doit servir de module de traitement des extragrammaticalités et s'intégrer au sein de systèmes qui ne sont pas conçus *a priori* pour le traitement de l'oral sans nécessiter des changements significatifs dans leurs architectures ou modules.

- **Précision :** par précision nous entendons la capacité du système à détecter et corriger uniquement les cas d'extragrammaticalité sans traiter les cas normaux même si ceux-ci présentent des similarités formelles avec des extragrammaticalités. Cette propriété est extrêmement importante pour un module de traitement des extragrammaticalité dans la mesure où des traitements erronés d'une extragrammaticalité peuvent conduire à des erreurs d'interprétation qui sont parfois plus graves que celles que peuvent créer les extragrammaticalités elles-mêmes.
- **Couverture :** le système doit être capable de traiter les différentes formes des extragrammaticalité quel que soit leur degré de complexité.
- **Simplicité :** le système doit être capable de traiter les extragrammaticalités avec le minimum de coût et les grammaires utilisées doivent être faciles à modifier.

Comme nous pouvons le deviner, certains de ces *requis* sont contradictoires. Par exemple, l'augmentation de la couverture de l'analyse augmente aussi les risques de surgénération.

1.2 Propriétés clés du système

Pour répondre aux différents requis, nous avons proposé un système dont les propriétés principales sont :

1.2.1 Emplacement dans le traitement

Afin de garantir l'indépendance totale à la fois du domaine d'application et du système au sein duquel le module de traitement des extragrammaticalités sera intégré, l'emplacement en tant que module de prétraitement semble la solution la plus appropriée. En effet, cela réduit considérablement l'interaction entre le module de traitement des extragrammaticalités et les autres modules du système. Par conséquent cela crée une autonomie des deux parties chacune par rapport à l'autre. Ainsi, le même module peut être utilisé dans différentes applications et avec des environnements logiciels et théoriques (quelque soit la nature de l'approche utilisée pour le module d'analyse). Par ailleurs, cela donne plus de liberté en ce qui concerne le choix des techniques de traitement puisque nous n'avons pas de contraintes externes à prendre en considération lors de la conception du module de traitement des extragrammaticalités. Finalement, comparé aux autres techniques de traitement des extragrammaticalités (en particulier aux techniques de post-traitement), le prétraitement permet de distinguer plus finement les types d'extragrammaticalités traités. En effet, les approches qui traitent les extragrammaticalités au cours de l'analyse syntaxique (avec une stratégie sélective par exemple) ou en postraitement (avec des règles sémantiques) ne permettent pas d'identifier le type de l'extragrammaticalité : avec les approches sélectives on perd toute trace de l'existence de

l'extragrammaticalité alors que les approches sémantiques ne permettent pas de distinguer les répétitions des auto-corrections. Par exemple, les segments *je voudrais j'aimerais* (auto-correction) et *je voudrais je voudrais* (répétition) ont une bonne chance d'avoir la même représentation sémantique et cela ne permet pas au module de post-traitement de savoir s'il s'agit d'un répétition ou d'une auto-correction.

Ainsi, notre approche ouvre la porte devant des expériences visant à tester l'utilité des informations relatives à l'existence des extragrammaticalités dans différents domaines applicatifs. Par exemple, dans le contexte d'un système de dialogue homme-machine, l'information sur l'existence d'une extragrammaticalité peut être prise en considération par le gestionnaire de dialogue afin de choisir la stratégie de dialogue la plus appropriée. Par ailleurs, l'identification des extragrammaticalités dans un système de traduction de la parole permet de générer l'équivalent de ces extragrammaticalités dans la langue cible et de donner, par conséquent, une dimension spontanée au dialogue en reflétant partiellement l'état psychologique des locuteurs exprimé par les extragrammaticalités.

Voici une présentation schématique de l'emplacement de notre module.

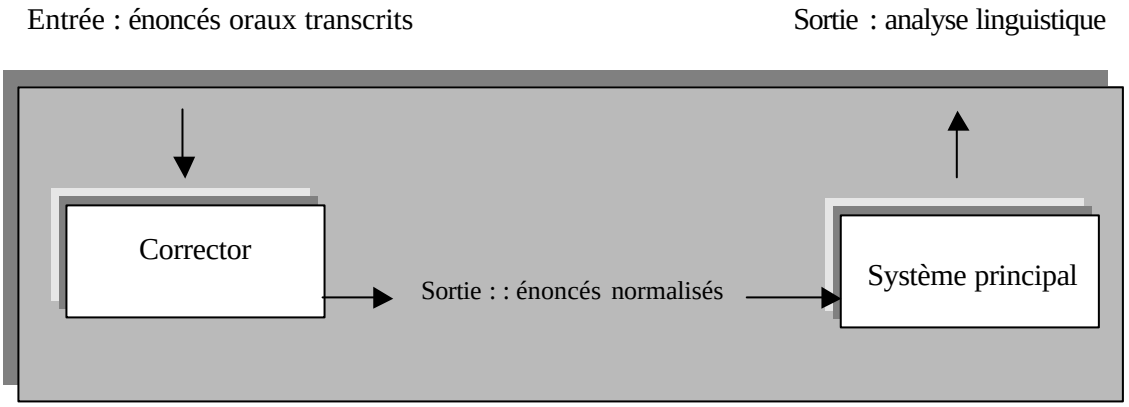


Figure 63. Schéma de l'emplacement du module de traitement des extragrammaticalités

Comme nous remarquons dans la figure précédente, l'interaction est limitée au traitement de la sortie du module de normalisation par le module suivant dans le système principal qui est considéré comme une boite noire.

1.2.2 L'architecture et les modules du système

Pour implanter les différentes fonctionnalités de Corrector nous avons adopté une architecture modulaire à base de Hub (Gestionnaire de système). Les motivations de notre choix ainsi qu'une discussion générale de l'architecture seront discutées plus loin.

Du point de vue du traitement, les phénomènes que nous avons obtenus lors de l'annotation du corpus, peuvent être classés en trois types :

1. Des phénomènes qui peuvent être traités avec l'information structurale uniquement représentée sous forme de patrons.

2. Des phénomènes qui peuvent être traités uniquement avec l'information morpho-syntaxique représentables avec une grammaire syntaxique superficielle ou la grammaire sémantique.
3. Des phénomènes nécessitant à la fois l'information syntaxique et l'information structurale. Ces phénomènes sont représentés avec des patrons mixtes.

Ainsi, nous avons proposé une architecture dans laquelle le traitement se fait par différents modules qui utilisent chacun l'une des trois techniques présentées ci-dessus. Cela se fait selon trois étapes principales :

1. Traitement lexical.
2. Traitement des Extragrammaticalités Supralexicales (ESLs) première passe.
3. Traitement des (ESLs) deuxième passe.

Le schéma général de cette architecture est présenté dans la figure suivante :

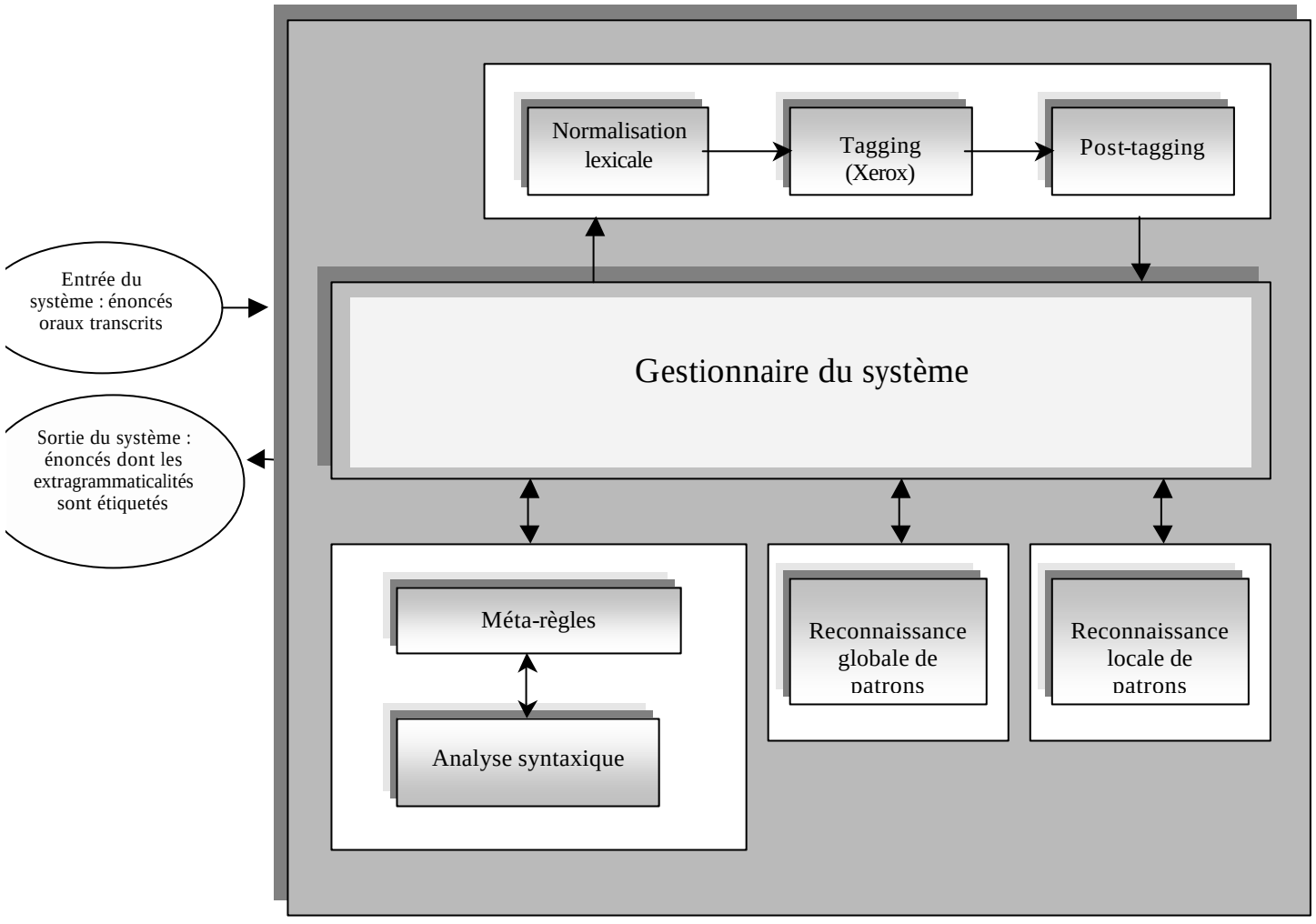


Figure 64. Architecture du système Corrector

Cette division modulaire renforce la portabilité du système. En effet, la répartition des tâches aux différents modules et en particulier l'indépendance du système vis à vis du *tagger* (le seul module externe).

2.1.1.63 Le gestionnaire du Système (GS)

Le GS est un *Hub* qui a une fonction purement logicielle et qui joue le rôle de corridor de l'information entre le reste des moduels du système (voir (Garlan and Shaw, 1993) pour plus d'informations sur les *Hubs* ainsi que leur différence avec les tableaux noirs). Ainsi, l'information au sein de ce module est propagée selon un ordre ascendant : du niveau lexical jusqu'au niveau syntaxique. Si nous considérons le GS comme un client qui appelle des fonctions diverses localisées au sein des modules du système (qui sont considérés comme des serveurs), le flux de l'information résultant de l'interaction peut être résumé dans le tableau suivant :

Entrée	Source	Destination	Traitement	Destination
Enoncé oral transcrit	Interface utilisateur	GS		
Enoncé oral transcrit	GS	Traitement lexical	Traitement des extragrammaticalités lexicales, normalisation de certains mots oraux et analyse morphologique	GS
Enoncé oral analysé morphologiquement et dont les phénomènes lexicaux sont normalisés	GS	Reconnaissance locale de patrons	Traitement des répétitions et des autocorrections dont l'étendue est limitée	GS
Enoncé dont les répétitions et les auto-corrections locales sont traitées	GS	Reconnaissance globale de patrons	Traitement des répétitions et des autocorrections dont l'étendue est large	GS
Enoncé dont les répétitions et les autocorrections sont traitées	GS		Traitement des faux-départs et des incomplétudes	GS
Enoncé dont les extragrammaticalités sont détectés et délimités	GS	Interface utilisateur		

Tableau 9. Le flux de l'information au sein du système Corrector

Outre la transmission de l'information entre les différents modules, le GS de corrector est aussi chargé d'adapter le format de sortie de chaque module au module suivant.

2.1.1.64 Traitement lexical

1.2.2.1.1 Normalisation lexicale

La fonction principale de ce module est la détection et le traitement des ELs. Outre l'évitement de certains problèmes que peuvent causer ces phénomènes aux analyseurs morphologique et syntaxique, certaines ELs peuvent causer des erreurs de détection au module de traitement des ESLs comme dans l'exemple suivant : *I'll uh I will*. On remarque dans cet exemple que l'amalgame cache la répétition de la séquence *I will*.

Ce module est basé sur une table de conversion contenant d'une part les différentes formes d'ELs rencontrées dans notre corpus et d'autre part les formes écrites de ces ELs.

1.2.2.1.2 Analyse morphologique (tagging et post-tagging)

La fonction principale de l'analyse morphologique est de fournir les parties du discours auxquelles appartiennent les mots de l'énoncé. Cette technique joue un rôle important dans le traitement des extragrammaticalités en particulier pour le traitement des autocorrections, faux-départs et incomplétudes.

La construction d'un tagger pour l'analyse morphologique étant une tâche qui dépasse largement nos moyens ainsi que les objectifs de notre travail, nous avons décidé d'utiliser un système déjà disponible.

Ainsi, sur ce point, notre travail s'est limité à choisir le tagger le plus adapté au traitement de l'oral parmi les systèmes disponibles (qui sont destinés au traitement de l'écrit) et à l'enrichir à l'aide de certaines fonctions de post-traitement afin de combler ses principaux lacunes par rapport à notre tâche. Notre choix est basé sur un test informel de quatre taggers disponibles sur Internet :

1. Le MBT tagger de l'équipe ILK (Induction of Linguistic Knowledge) à l'université de Tilburg[35].
2. Le tagger du groupe MLTT du laboratoire de Xerox à Grenoble.
3. QuickTag de Cogilex (entreprise canadienne basée à Montréal).
4. CLAWS tagger du groupe UCREl de l'université de Lancaster.

Le choix a été fait sur la base d'un test informel d'une vingtaine d'énoncés qui contiennent des extragrammaticalités de différents types que nous avons extraits du *Trains corpus*. Deux principaux critères ont été retenus pour l'évaluation des systèmes :

a. L'adaptation de la sortie : elle porte essentiellement sur la finesse de l'analyse et son adaptation à nos besoins. Par exemple, les systèmes MBT, QuickTag et CLAWS sont dotés d'une fonction particulière de traitement des mots inconnus qui associe des catégories morphologiques induites du contexte (adjectif, nom, etc.) aux mots inconnus. Cela nous

[35] Ce tagger peut être testé à l'URL suivant : http://ilk.kub.nl/

empêche de détecter les mots incomplets (qui sont inconnus par le système) et de les traiter correctement. Par contre, le tagger de Xerox associe à ces mots une catégorie morphologique tout en indiquant que ces mots ne font pas partie de son lexique. Par exemple, ce système associe la catégorie : +guessed+ADJ à un mot inconnu, qui selon son contexte, peut être considéré comme un adjectif.

b. La qualité des résultats : nous avons accordé une attention particulière à la qualité d'analyse en cas d'extragrammaticalité, pour choisir l'analyseur le plus robuste.

Le résultat de ce test a été clairement en faveur du tagger de Xerox, d'une part, parce que sa sortie, comme nous avons vu, est plus adaptée que les autres systèmes et d'autre part parce que nos tests ont montré qu'il est plus le plus robuste aux extragrammaticalités.

Malgré notre effort de choisir le système le plus adapté, le tagger de Xerox étant un outil *généraliste*, il est normal que des incomplétudes ou des inadaptations partielles de son fonctionnement soient observées par rapport à notre tâche spécifique. Deux aspects relatifs au tagging ont été observés et traités :

1. **Le manque de finesse :** dans certains cas, le traitement d'une extragrammaticalité nécessite des informations qui vont au-delà de la simple catégorie morphologique. Par exemple, dans certains cas, nous avons besoin de savoir si un pronom est sujet ou objet pour juger s'il est un complément d'un syntagme verbal précédent SV_p ou d'un syntagme verbal suivant SV_s afin de décider si la phrase constituée par SV_p est complète ou non. Ainsi, nous avons décidé d'augmenter la sortie du tagger afin de l'adapter à nos besoins. Pour simplifier le traitement, cet enrichissement n'est fait que dans les cas où l'on en a besoin. En effet, il ne s'agit pas d'un module de post-tagging, mais plutôt d'une base lexicale à laquelle le système fait appel au cours de l'analyse syntaxique lorsqu'il y a un segment dont le traitement nécessite la vérification affinée des propriétés morpho-syntaxiques d'un de ses mots. La base lexicale utilisée consiste en la série des mots à enrichir associés à leurs nouvelles catégories. Il s'agit essentiellement des verbes transitifs observés dans le corpus ainsi que des pronoms personnels sujet.
2. **Des erreurs relatives à l'application :** il s'agit généralement d'erreurs prévisibles et répétitives d'analyse de mots propres au domaine de notre corpus. Par exemple, les mots : Corning et Coring qui sont des noms de lieux dans notre corpus sont considérés par le tagger comme étant des *participes présents* ou des *adjectifs* dont les racines sont respectivement : *corn*, et *core* (que nous n'avons pas observé dans notre corpus). Ce genre de cas est corrigé directement avec un module de post-tagging. Il s'agit d'une simple table de conversion qui contient d'une part, les mots que le système analyse incorrectement de manière systématique et d'autre part, les versions correctes de leur analyse.
3. **Des erreurs dues à des raisons diverses :** ce sont des erreurs occasionnelles dues au tagger lui-même ou aux extragrammaticalités dans notre corpus. Ces erreurs sont pratiquement

impossibles à corriger avec des post-traitements. Pour éviter ce problème, nous avons décidé de réduire au maximum l'utilisation des informations morphologiques. Ainsi, comme nous allons voir en détail dans les paragraphes suivants, le traitement d'une bonne partie des extragrammaticalités supralexicales se fait à l'aide de patrons ne nécessitant que des informations structurales. Par ailleurs, nous avons introduit les grammaires sémantiques pour le traitement de certaines zones d'édition, ce qui réduit le besoin des catégories morphologiques et finalement, nous avons opté pour des règles syntaxiques souples (analyse partielle par segments) qui nécessitent le recours à un contexte assez large afin de réduire l'effet des erreurs locales de tagging.

Le lien avec le serveur de Xerox où se trouve le tagger et notre système se fait avec un script qui envoie les énoncés pré-normalisés au tagger, récupère la sortie du tagging et la formate de manière à la rendre adaptée au module suivant. Ce script est une version que nous avons adaptée du code de notre collègue José Rouillard qui est utilisée pour son système Halpin (Rouillard, 2000).

2.1.1.65 La reconnaissance de patrons

La reconnaissance de patron est un dispositif économique et facilement généralisable et portable d'une application à une autre voire d'une langue à une autre, dans certains cas. De plus, il est facile à intégrer avec d'autres techniques de traitement.

1.2.2.1.3 Présentation informelle de notre approche

Comme nous avons vu dans la partie théorique, la différence principale entre œtte technique et l'analyse grammaticale normale est que dans ce cas, nous avons des informations structurales basées sur l'identité des mots à côté des informations morphologiques qui peuvent être présentes dans certains patrons. Ainsi, deux types de patrons ont été utilisés :

- **Des patrons simples :** il s'agit des patrons basés uniquement sur les informations structurales comme le patron : M1 M2 M1 M2, où l'on a besoin de vérifier uniquement l'identité du mot et son emplacement dans la chaîne.
- **Des patrons hybrides :** ce sont des patrons qui combinent l'information structurale à la morphologie ou même à la grammaire sémantique. L'information morphologique consiste en l'enrichissement des patrons de certains éléments dont le traitement se fait non pas en considérant leur identité mais plutôt avec leur catégorie morpho-syntaxique et sa relation avec celle d'autres éléments. Pour mettre au clair ce point, examinons le patron suivant : M1 M2 R1 M1 M2 R1'. Dans ce patron, les éléments répétés (représentés par *M*) sont analysés en considérant leur identité et leur emplacement dans la chaîne. Par contre, les mots représentés par un *R* (qui correspondent à des remplacements) sont traités selon leurs catégories morphologiques respectives. En général, il s'agit de deux mots différents dont les catégories sont identiques ou assez proches fonctionnellement, comme : un cardinal et un déterminant. Sachant que l'ordre d'apparition de ces éléments dans l'énoncé est aussi pris en considération.

Les informations sémantiques sont intégrées dans les patrons afin de représenter la zone d'édition. Par exemple, dans le patron : M1 Ed M1 l'élément Ed peut correspondre à une règle d'une grammaire sémantique. Cette règle peut être : Ed → Verb_wait det Moment_word (wait a moment).

Pour l'implantation du module de reconnaissance des patrons, la première phase de notre travail a consisté à étendre et généraliser certains patrons obtenus lors de la phase d'analyse théorique. Voici quelques exemples de ce processus :

- La transition interdite entre deux catégories identiques utilisée comme critère pour la détection de l'autocorrection a été généralisée à toutes les catégories avec certaines exceptions comme pour les cardinaux.
- Nous avons ajouté des patrons avec des zones d'édition pour tous les patrons sans zone d'édition et pour lesquels nous n'avons pas observé un équivalent avec zone d'édition. Nous avons aussi fait l'opération inverse pour les patrons observés uniquement avec des zones d'édition. Par exemple, le patron R1 M1 M2 R1'M1M2 a été observé uniquement sans zone d'édition mais. Ainsi, nous avons ajouté la version avec une zone d'édition : R1M1M2 Ed R1'M1M2 a l'ensemble de nos patrons.
- Nous avons étendu certains patrons analogiquement. Par exemple, le patron M1M2M3M4 M2M1M3M4 (qui correspond à une autocorrection par inversion) a été généralisé à l'autocorrection avec répétition de trois mots et cinq mots.

Ainsi, nous avons augmenté le nombre de nos patrons d'environ 22,9% et nous avons obtenu ainsi un nombre total de 61 patrons (sans considérer les variations de la zone d'édition).

Nous avons implanté les patrons obtenus avec un mécanisme général de parcours descendants (que nous allons présenter de manière détaillée plus loin). Après avoir implanté ces patrons, nous étions devant le problème de choisir lequel des patrons activer selon les contextes afin d'éviter les surgénérations.

1.2.2.1.4 Le contrôle de l'application des patrons

Certains de ces problèmes sont automatiquement résolus grâce aux propriétés internes des patrons alors que certains d'autres ont nécessité l'implantation d'algorithmes spécifiques ou l'adoption d'une stratégie d'analyse particulière. Les principaux moyens de réduction de surgénération sont présentés dans les points suivants :

1. **Les contraintes internes des patrons :** le principe de base qui contrôle l'intervention d'un patron quelconque dans le traitement d'une extragrammaticalité est ses propres contraintes. Pour mettre au clair ce principe, prenons l'exemple d'une répétition simple d'un seul mot :

   yeah yeah (**99**)

   Pour traiter cette répétition, le système cherche d'abord le patron correspondant parmi tous les patrons possibles. Cette recherche se fait selon les deux étapes suivantes :

i. **Elimination des patrons dont la taille n'est pas correspondante :** cette élimination se fait essentiellement sur la base de la taille de la fenêtre correspondant au patron. Par exemple, les patrons M1EM1 ou M1M2M3 M1M2M3 sont automatiquement éliminés puisqu'ils nécessitent des extragrammaticalités dont l'étendue en terme de mots est plus grande que le segment en cours d'analyse.

ii. **Vérification du patron dont la taille est correspondante :** la deuxième étape consiste à vérifier si les contraintes du patron dont la taille est correspondante sont satisfaites dans le segment en cours d'analyse : si oui, alors le patron en question est associé à ce segment. Sinon, ce segment est considéré comme un segment grammatical.

2. **L'ordonnancement des patrons :** lorsque les contraintes de plusieurs patrons peuvent être satisfaites par le même segment alors on parle de conflit de patrons. Le conflit existe souvent entre des patrons de répétition et des patrons d'autocorrection. Prenons l'exemple suivant :

I want I want (100)

Deux patrons sont applicables pour le traitement de ce phénomène :

- M1 R1 M1 R1 (Autocorrection)
- M1 M2 M1 M2 (Répétition)

Pour résoudre l'ambiguïté, le système procède de manière déterministe, c'est-à-dire, il prend la première solution satisfaisante et se désintéresse du reste. Malgré ses avantages en terme de rapidité et simplicité de traitement, cette approche peut conduire à l'erreur si les solutions ne sont pas bien ordonnées. Ainsi, dans notre exemple, si le système examine le patron de l'autocorrection d'abord, il décidera que le segment en cours d'analyse est une autocorrection puisque d'une part, les deux premiers mots de chaque côté de l'extragrammaticalité (les deux *I*) sont identiques et d'autre part, les deux mots *Want* ont la même catégorie morphologique. Dans ce cas, le patron de répétition ne sera pas examiné et le système décidera incorrectement qu'il s'agit d'une autocorrection. Pour éviter ce problème, nous avons ordonné les patrons du plus contraignant au moins contraignant (autrement dit du moins sur-génératif au plus sur génératif). Ainsi, dans notre exemple, nous avons placé le patron de répétition avant le patron d'autocorrection ce qui permet d'éviter la surgénération dans les deux sens puisque les patrons de répétition n'acceptent pas l'autocorrection et par conséquent le système est obligé de vérifier le patron d'autocorrection et donne l'analyse correcte.

3. **Les patrons de contrôle :** comme nous avons vu dans notre étude théorique, dans certains cas, des expressions linguistiques particulières comme ***to*** *go* ***to***, ***as*** *soon* ***as*** ainsi que certains phénomènes comme le comptage *one two, three,* ont la forme d'une extragrammaticalité et exigent un traitement particulier afin qu'ils ne soient pas corrigés par erreur. Ainsi, nous avons recensé dans notre corpus d'apprentissage 16 formes de surgénération qui ont été représentée avec des patrons et des règles de contrôle qui ont été privilégié dans l'ordonnance afin

d'empêcher le système de reconnaître les segments qui satisfont leurs contraintes comme extragrammaticaux.

4. La double passe : il s'agit d'une solution que nous avons proposée pour le traitement des extragrammaticalités imbriquées. Les détails de cette solution seront présentés plus loin dans le paragraphe dédié au traitement des extragrammaticalités imbriquées.

1.2.2.1.5 Présentation formelle de l'algorithme de reconnaissance des patrons

Pour présenter l'algorithme de reconnaissance de patrons, nous avons adopté le cadre de l'analyse déductive (parsing as deduction) présenté par (Shieber *et al.*, 1995). L'adoption de ce cadre est justifiée par plusieurs raisons :

1. Il s'agit d'un cadre général qui permet de représenter des algorithmes de types variés. En effet, ce cadre a été choisi pour la présentation de différents algorithmes de types divers (voir (Shabes et Waters, 1995), (Lopez, 1999a), (Goodman, 1999) comme exemple de travaux qui ont adopté ce cadre).

2. L'aspect formel de ce cadre nous permet de présenter et discuter les différentes propriétés de nos algorithmes.

3. L'utilisation de ce cadre pour présenter les différents algorithmes d'analyse grammaticale que nous avons implanté dans ce travail nous permet de les comparer et de montrer leur complémentarité.

I. **Définition de la grammaire utilisée pour la reconnaissance des patrons :** Soit $G = (S, N, \mathring{a}, R)$ où :

1. S est l'ensemble des non-terminaux distingués de la grammaire. Contrairement aux grammaires classiques où il existe un seul symbole distingué dans la grammaire, notre grammaire contient un ensemble de non-terminaux correspondant chacun à un îlot autorisé par la grammaire.
2. N est le vocabulaire non-terminal de la grammaire.
3. $\mathring{a}$ est l'ensemble des terminaux. Ainsi le vocabulaire $V = \mathring{a} \cup N$.
4. R est un ensemble de règles de réécriture dont le schéma est le suivant : $A \rightarrow \alpha$ où $A \in N$ et $\alpha \in V^*$.

II. **La notation :** soit la chaîne de mots à analyser : $W = w_1 \ldots. w_n$, l'unité de base que nous allons adopter pour la présentation des opérations de notre algorithme a la forme suivante $[\cdot b, j]$, où $0 \leq j \leq n$. Cette unité signifie que la phrase du langage peut être obtenue par la sous-chaîne de w : $w_1 .. w_j$ (w_1 et w_j inclus) suivi par la chaîne de symboles b. En d'autres termes $S \Rightarrow^{*}$ [36] $w_1 .. w_j b$.

[36] Le symbole $\Rightarrow^{*}$ est utilisé pour désigner les dérivations réflexives.

Par ailleurs, notons que le point dans l'unité de base est utilisé pour séparer la partie qui a été analysée de celle qui ne l'a pas été encore.

III. L'algorithme : avant de présenter notre algorithme, nous allons procéder à une formalisation des patrons que nous avons utilisés. En effet, la définition formelle des patrons est basée sur l'idée de symétrie entre les deux segments répétés. Ainsi nous avons distingué entre trois schéma de partons :

1. **Les répétitions simples :** la définition des répétitions simples est basée sur la symétrie des éléments impliqués dans une répétition ainsi que leur identité. Soit le prédicat unifiy(Arg_1, Arg_2), qui est vrai si et seulement si Arg_1 s'unifie avec Arg_2 et soit la chaîne de mots $W = w_1 \ldots w_n$, une sous-chaîne de W : $W_R = w_i .. w_j$ (où $1 \leq i < n$ et $1 < j \leq n$) est jugée comme étant une répétition si et seulement si $\forall\ w_x$ où $i \leq X < (j-i)-1$ alors unif(w_x, $w_{(j-i+1)/2+x}$).
2. **Les répétitions avec zone d'édition :** comme la symétrie totale entre les éléments d'une répétition avec zone d'édition n'est plus existante (à cause de la zone d'édition qui apparaît au milieu de la répétition), nous allons procéder d'une manière légèrement différente pour définir ces phénomènes. Ainsi, si nous prenons une sous-chaîne de W, $W_{RE} = w_i .. w_j\ e_1 \ldots e_n\ w_i' .. w_j'$ cette sous-chaîne est considérée comme une répétition avec une zone d'édition si et seulement si $\forall\ w_x$ où $i \leq X < j$, alors unif(w_x, w'_x) $\wedge$ Ed[37] $\rightarrow$ $e_1 \ldots e_n$.
3. **Les auto-corrections :** la différence principale entre les patrons des auto-corrections est que tous les mots ne sont pas identiques : certains mots sont répétés alors que certains d'autres sont remplacés (en général il s'agit d'un seul mot). Voici les définitions des prédicats et unités nécessaires pour la présentation des règles d'inférence pour le traitement des auto-corrections :

- Soit le prédicat, *replace*(C_1, C_2) qui est vrai si la catégorie morphologique C_2 est acceptée comme une catégorie qui peut remplacer C_1 (les valeurs de C_1 et C_2 sont stockées dans le systèmes sur la base des observations des auto-corrections dans le corpus),
- Soit une sous-chaîne de W, $w_{ac} = w_i .. w_j$ (où $1 \leq i < n$ et $1 < j \leq n$) et
- Soit le prédicat *location*(C_{wx}, x) qui est vrai si et seulement si x ($i \leq x \leq j$) correspond à la position du mot w_x dont la catégorie morphologique est C_{wx} dans la sous-chaîne w_{ac}. La valeur de x est prédéfinie sur la base des observations des auto-corrections dans le corpus. Par exemple, dans le patron : M3R1 M3R1', le prédicat *location*(*R1*, *4*) permet de préciser la location de l'élément remplacé à partir du premier mot du patron.

[37] ED est un non-terminal qui couvre une zone d'édition acceptable par la grammaire.

Ainsi, w_{ac} est considérée comme une auto-correction si et seulement si $\forall\ w_x$ tel que *location*(C_{wx}, *x)* alors *replace*(w_x, $w_{(j-i+1)/2+x}$) et $\forall\ w_y$ où $y \neq x$ et $i \pounds X < (j-i)-1$ alors unify(w_x, $w_{(j-i+1)/2+x}$).

Pour des raisons de concision, nous allons donner l'algorithme avec les schémas des patrons seulement (nous n'allons pas énumérer tous les patrons que nous avons utilisés). Ainsi, l'algorithme d'analyse a la forme suivante :

Axiome : $[\cdot\ s_x, 0]$

Objectif : $[\cdot, n]$

Scan :

$$\frac{[\cdot\ w_{i+1}...w_j b,i]}{[\cdot\ b, j]}$$

$\forall\ w_x$ où $i+1 \pounds x < (j-(i+1))-1$ alors *unify*(w_x, $w_{(j-i)/2+x}$)

$$\frac{[\cdot\ w_{i+1} .. w_j e_1 \ldots e_n w_i' .. w_j' b,i]}{[\cdot\ b, j']}$$

$\forall\ w_x$ où $i+1 \pounds x < j$, alors *unify*($w_x, w_{x'}$) $\wedge$ $Ed \rightarrow e_1 \ldots e_n$

$$\frac{[\cdot\ w_{i+1} .. w_j\ b,i]}{[\cdot\ b, j+1]}$$

($\forall\ w_x$ tel que *location*(C_{wx}, *x)* alors *replace*(w_x, $w_{(j-i+1)/2+x}$))

$\wedge$

($\forall\ w_y$ où $(y \neq x) \wedge (i+1 \pounds y < (j-i)-1)$ alors *unify*(w_y, $w_{(j-i)/2+y}$))

$$\frac{[\cdot\ w_{j+1} b,j]}{[\cdot\ b, j+1]}$$

Prédiction :

$$\frac{[\cdot\ Bb,j]}{[\cdot\ gb,j]}$$

$B \rightarrow g$

Figure 65. L'algorithme de reconnaissance de patrons

Comme tout algorithme descendant, notre reconnaisseur de patron commence en émettant l'hypothèse que l'entrée en cours d'analyse peut-être analysée par l'un des non-terminaux distingués de la grammaire. Ainsi, il suppose l'élément $[\cdot\ s_x, 0]$ (qui signifie que l'énoncé peut être analysé avec $s_x \in S$) et essaie de prouver $[\cdot, n]$ (qui veut dire que tous les éléments de l'entrée ont été analysés avec s_x). Ensuite, l'algorithme applique les patrons de manière

descendante en commençant par les patrons les plus restrictifs (les patrons correspondant à des répétitions) pour arriver aux patrons les moins restrictifs (les patrons des faux-départs). Finalement, après avoir tenté les différents patrons qui peuvent s'appliquer à l'entrée, notons que l'algorithme est capable de traiter les mots qui ne font pas partie d'un patron en utilisant la règle suivante :

$$\frac{[\cdot\ w_{j+1}\, b, j]}{[\cdot\ b, j+1]}$$

Après avoir utilisé cette règle, si la totalité des mots de l'énoncé n'ont pas été consommés, l'algorithme commence un nouveau cycle jusqu'à la fin de l'entrée. Notons, que si un s_x est complètement satisfait et si l'entrée n'est pas complètement analysée, l'algorithme tente de nouveaux s_x (les s_x sont classés selon leur priorité). Finalement, il n'est probablement pas inutile de mentionner que chaque sous-groupe de patron correspond à un seul s_x. Cela évite de parcourir tous les patrons à chaque émission d'une hypothèse.

4. **Discussion de l'algorithme :** la complexité de l'algorithme est équivalente au nombre des variables libres dans la règles la plus complexe, c'est-à-dire que la complexité est équivalente au patron le plus long $O(n^8)$. Cette complexité peut être réduite à $O(n^{(x/2)+1})$ où x est le nombre des mots du patron le plus long. Cette réduction peut être faite en utilisant des techniques tabulaires qui prennent une fenêtre de $(x/2)+1$ mots : compare le premier mot de cette sous-chaîne avec son avant-dernier et puis avance d'un mot et répéter la même procédure jusqu'à la couverture de la totalité des mots impliqués dans l'extragrammaticalité.

   Cette technique malgré son intérêt théoriquement n'est pas nécessaire pour améliorer la performance pratique de l'algorithme. En effet, l'algorithme de reconnaissance de patrons a des performances proches du temps réels et ne nécessite pas une amélioration majeure. Cette performance est justifiée par les raisons suivantes :

   i. Le nombre des patrons utilisés dans l'application est assez limité. En effet, comme nous avons vu, le nombre total des patrons utilisés (avec les ajouts que nous avons effectués) est 61 patrons. Cela réduit considérablement l'espace de recherche de l'algorithme et conséquent augmente sa rapidité.

   ii. Nous avons organisé les patrons de manière à éviter de parcourir tous les patrons à chaque émission.

   iii. Comme montré dans (Abney, 1995), l'application de plusieurs passes dans l'analyse contribue à augmenter la rapidité du traitement étant donné que la complexité due à l'interaction des niveaux d'analyse (dans notre cas les patrons locaux et les patrons globaux) est réduite.

2.1.1.66 L'étiquetage syntaxique par Réseaux de Transition Récursifs RTRs

1.2.2.1.6 La tâche du module d'étiquetage syntaxique

Deux étapes séparées sont nécessaires pour le traitement de ces phénomènes : la détection et la délimitation. A son tour, la détection est basée sur deux facteurs :

a- Localisation du centre du faux-départ ou de l'incomplétude.

b- Délimitation de l'étendue de l'extragrammaticalité localisée en détectant ses frontières. Cette délimitation nécessite non seulement la détection du segment extragrammatical mais aussi la précision de tous les segments qui dépendent de lui ou des quels il dépend. La fonction principale de cette précision étant la délimitation de la zone à corriger. Prenons l'énoncé suivant :

We need a shorter route **from** we need to um manage to get the bananas to Dansville more quickly <sil> um (Utt42, d93-14.3) (101)

Nous remarquons que dans l'énoncé précédent, bien que le constituant prépositionnel *from* est le seul à être incomplet (puisqu'il nécessite d'être suivi par un syntagme nominal ou un pronom objet qui ne sont pas présents), il nous faut délimiter tous les segments qui sont directement liés à lui. En général, il faut marquer le prédicat syntaxique dont dépend le constituant incomplet aussi bien que tous les autres constituants qui dépendent de ce prédicat. Ainsi, dans notre exemple, il faut marquer le prédicat verbal *need* ainsi que les deux constituants qui dépendent de lui *we* et *a shorter route* ainsi que le constituant incomplet *from.* Dans certains cas, notamment lorsque le faux-départ est situé au milieu de l'énoncé, la tâche de délimitation des constituants qui dépendent du prédicat dominant le segment mal formé s'avère plus difficile. Cela nécessite la combinaison des sources d'informations syntaxiques et supra-syntaxiques pour délimiter ses frontières de début et de fin.

1.2.2.1.7 Les Réseaux de Transition Récursifs RTRs

Les RTRs sont une version étendue des FSA (voir (Woods, 1970) pour une présentation de ce dispositif). Tout comme les FSAs (Finite State Machines), ils sont composés d'une série d'états et de transitions. Il s'agit d'un graphe étiqueté dont chaque étiquette correspond à une catégorie (lexicale, syntaxique ou conceptuelle) la transition d'un état à un autre est subordonnée par la réussite de l'unification entre d'une part l'étiquette de l'arc et d'autre part le mot ou (le sous réseau) courant. Ainsi, un état dans un RTR consiste en quatre éléments :

1. **Le nœud/réseau :** cet élément fournit de l'information sur la location du traitement.
2. **Le reste de la phrase :** indique la partie de la phrase qui n'est pas encore analysée.
3. **Les nœuds en attente :** les nœuds dans le réseau en cours qui ne sont pas encore traversés.
4. **L'analyse :** il s'agit de l'analyse associée à la partie traitée de la phrase d'entrée.

Trois actions sont possibles lorsque l'analyseur est dans un état particulier selon la nature de cet état :

1. **L'étiquette est une catégorie syntagmatique (sous-réseau) :** mettre le nœud en cours dans la pile d'attente et créer un nouveau constituant pour une nouvelle catégorie.
2. **L'étiquette est une catégorie lexicale :** vérifie l'identité de ce mot et ajouter ce mot ainsi que sa catégorie au constituant en cours.
3. **Le constituent est complet :** prendre le nœud en attente de la pile et intégrer le constituent en cours dans un constituent de niveau supérieur.

De manière plus formelle, une chaîne *S* composée d'un ensemble de sous-chaînes s_1 .. s_k tel que $S = s_1$ .. s_k cette chaîne est reconnue en tant que *X* par un réseau *N* si et seulement si :

4. *X* est l'étiquette d'un état initial *x* et d'un état final *y* (où *x* et *y* correspondent respectivement à 1 et *k*) et
5. Il existe un chemin (une chaîne d'étiquettes) $l_1 \ldots l_k$ accepté par *N* (vu comme un réseau de transition non-récursif) et avec *x* comme état initial et
6. Pour chaque s_i (où $k \geq i \geq 1$) soit $s_i = l_i$ (dans ce cas s_i correspond à un mot) ou s_i est reconnu comme un sous-réseau l_i.

Ainsi, contrairement aux grammaires syntagmatiques qui consistent en séries linéaires de symboles, les RTRs constituent un treillis de symboles. Afin de rendre compte des composantes du treillis de symboles crée par un RTR, nous avons adopté la notation suivante :

Notation	Type d'arc
$?_{SR}$	Début de la Séquence d'une Règle
$?^{-1}_{SR}$	Fin de la Séquence d'une Règle
$?_{RA}$	Début des alternatives à une règle
$?^{-1}_{RA}$	Fin des alternatives à une règle
$?_{TAV}$	Transition avant vide
$?_{TArV}$	Transition arrière vide

Tableau 10. Les étiquettes adoptées pour l'annotation des RTRs

Voici un exemple d'un réseau de transition présenté avec la notation que nous avons adoptée :

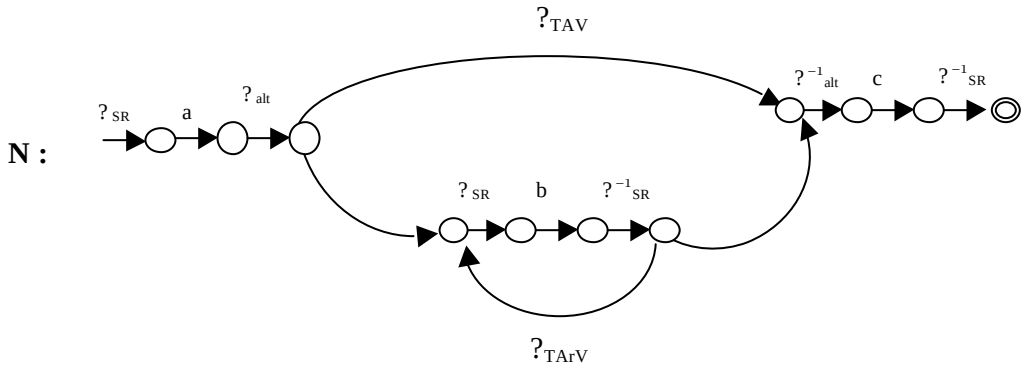


Figure 66. Exemple d'un réseau de transition récursif

Ce réseau permet de reconnaître des chaînes comme : *a c* (la transition avant vide permet de ne pas considérer *b*), *a b c*, *a b b c* (la transition arrière vide permet d'accepter un nombre infini de *b*), *a b b b c*, etc.

Bien qu'ils soient équivalents aux CFGs, les RTRs présentent plusieurs avantages par rapport à elles :

1. Les RTRs sont plus compacts et plus efficaces que les règles syntagmatiques classiques. En effet, un RTR peut couvrir plusieurs règles. Pour mettre au clair cette idée examinons la petite grammaire suivante au format DCG[38] (pour la clarté de l'exposé, nous avons omis les règles dont la partie droite est un terminal) :

[38] Definite Clause Grammar.

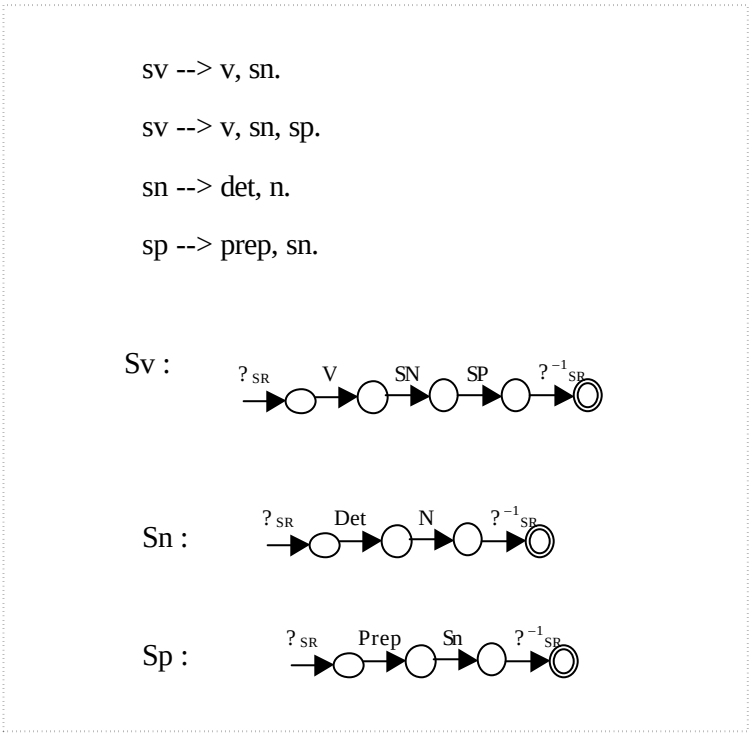


Figure 67. Une mini-grammaire CFG et les RTRs correspondants

La première remarque que nous pouvons faire à propos de cette mini-grammaire et les RTRs équivalents est que les règles correspondantes au *Sv* sont contractées au sein d'un seul réseau. Outre l'avantage de cette concision de représentation, le traitement avec un RTR est plus efficace qu'avec la grammaire syntagmatique. Supposons que nous voulons analyser l'énoncé : *the dog chased the cat near the elephant.* Avec un algorithme descendant qui utilise la grammaire syntagmatique, tout d'abord le système essaye la première règle dans la partie gauche est (V Sn) et comme la totalité de l'énoncé n'est pas encore analysée, il essaye la deuxième alternative de *Sv* qui comprend un *Sp* aussi. Le problème est qu'avec la deuxième tentative, le système doit recommencer à nouveau l'analyse du verbe et du sn qui ont été correctement analysés la première fois. Comme les deux alternatives du *Sv* sont codées avec un seul réseau, les deux éléments communs aux deux règles de *Sv* sont gardées lorsque le système essaye de vérifier les éléments non-partagés entre ces deux formes. Cette propriété rend les RTRs comparables aux algorithmes tabulaires (qui conservent un historique des analyses partielles). Cependant, une différence notable entre les RTRs et les algorithmes tabulaires mérite d'être citée. En effet, les tables d'analyse dans les algorithmes tabulaires sont crée en-ligne (au cours de l'analyse) alors que dans les RTRs, le graphe correspondant à la grammaire est crée hors-ligne lors de la compilation de la grammaire.

2. La possibilité d'exprimer des répétitions infinies : il est assez facile d'exprimer la répétition infinie d'un élément quelconque dans la grammaire. Cette propriété est particulièrement utile

pour l'implantation de la stratégie sélective aussi bien que la division de la grammaire en sous-grammaires selon le domaine (nous allons voir ces deux aspects avec plus de détails plus loin).

Les RTRs ont été un paradigme très populaire dans les années soixante-dix et quatre-vingt pour des tâches d'analyse syntaxique. Ils ont été récemment utilisés pour l'implantation de grammaires sémantiques pour le traitement de dialogues oraux (voir entre autres (Ward, 1991), (Mayfield, *et al.*, 1995), (Gavaldà, 2000)).

Ainsi, nous avons utilisé des réseaux qui représentent les principales zones observées dans notre étude théorique. Pour augmenter l'efficacité de traitement, les RTRs ont été équipés d'un filtre lexical qui permet d'établir une première vérification de la validité de l'hypothèse émise selon le mode descendant avec le segment de l'énoncé en cours de l'analyse. Par ailleurs, les RTRs utilisés ont été enrichis par une stratégie sélective qui permet au système de détecter les zones qu'il peut analyser et d'ignorer le reste. Cette stratégie est similaire à celle utilisée dans le système OASIS que nous allons présenter dans le deuxième chapitre de cette partie où les différentes propriétés de cette stratégie sélective seront présentées en détail.

1.2.2.1.8 Présentation formelle de la version des RTRs que nous avons implantée

1. **Notation et prédicats de base :** soit la chaîne de mots à analyser : $W = w_1 \ldots w_n$, l'unité de base que nous allons adopter pour la présentation des opérations de notre algorithme a la forme suivante $[N, \cdot Cb, j]$, où $0 \leq j \leq n$. Cette unité signifie, que pour traverser le réseau N, il faut traverser le nœud étiqueté par la catégorie C et qui commence à l'état j. Notons que b (qui est une chaîne de symboles) représente la partie de l'énoncé qui n'a pas encore été consommée par le réseau N. Cette partie est implantée sous forme d'une pile dont on consomme des mots incrémentalement avec le progrès de l'analyse. Par ailleurs, notons que le point $(\cdot)$ dans l'unité de base est utilisé pour séparer la partie qui a été analysée de celle qui ne l'a pas été encore.

   Voici les prédicats que nous avons utilisés dans notre présentation des RTRs :

   - Le prédicat *coin_gauche(N, W)* qui est vrai si et seulement si le premier mot de W (ou son coin gauche) w_1 fait partie des mots avec lesquels N peut commencer. Notons, que cette information est obtenue à partir d'une table créée automatiquement lors de la compilation de la grammaire et qui contient la liste des mots à partir desquels un réseau peut commencer.
   - Le prédicat *lex(C, M)* est vrai si et seulement si C est définie dans le lexique du système comme étant la catégorie grammaticale du mot M.
   - Le prédicat *arc(a, b, L, N)* est vrai si et seulement s'il existe dans le graphe du réseau N un chemin qui commence par l'état a et qui se termine par l'état b.
   - Le prédicat *initial(N, x)* qui est vrai si et seulement si x est l'état initial du réseau N.

- Le prédicat *final(N, y)* est vrai si et seulement si *y* est l'état final du réseau *N*. Notons que $y \geq x$.
- Finalement le prédicat traverse(C, w_{j+1},*b, j*) est vrai si et seulemen si : (lex(C, w_{j+1})) $\vee$

  (initial(C, x) $\wedge$

  coin_gauche(C, W) $\wedge$
  recognize(C, x, w_{j+1}, *b*))

2. **L'algorithme :** voici une présentation d'une version simplifiée des RTRs que nous avons utilisés dans Corrector[39] :

Axiome :	[*N*, · *W*, *0*]	initial(N, 0)
Objectif :	[*N*, ·, *n*]	final(N, n)

$$\text{recognize}\quad \frac{[N, \cdot\, w_{j+1} b, j]}{[N, \cdot\, b, l]} \qquad \text{arc}(j, l, L, N) \wedge \text{traverse}(L, w_j, 1b)$$

Figure 68. Présentation formelle des RTRs que nous avons utilisés

L'axiome de l'algorithme veut-dire que le réseau *N* dont les arcs commencent à l'état 0 et se termine à l'état $n \geq 0$ permet d'analyser la chaîne de mots *W*.

La formule de l'objectif (ou la clause d'arrêt) veut-dire que le réseau *N* est considéré comme satisfait si tous les mots de l'entrée sont consommés par ce réseau (la pile des mots à analyser est vide) et si le réseau arrive à son état final. Finalement, l'opération recognize permet de passer d'un état à un autre, s'il existe dans le graphe de *N* un arc qui lit ces deux états et si l'élément en cours d'analyse peut satisfaire le prédicat traverse : il doit être analysé soit comme un item lexical soit comme un sous-réseau.

3. Discussion de l'algorithme :
   a. L'aspect déterministe : l'aspect déterministe consiste à encoder dans la grammaire un ensemble de préférences pour la résolution des conflits entre les règles de la grammaire (qui sont causés généralement par les ambiguïtés). Différents algorithmes déterministes ont été implantés pour des applications d'analyse grammaticale (Hindle, 1983), (Sabah et Rady, 1983), (Briscoe, 1987). Ainsi, dans nos grammaires deux principes généraux ont été respectés afin de résoudre les ambiguïtés :

[39] Pour la clareté de l'exposé, nous avons omis l'opération de traverse spéciale pour les arcs faculatatifs dans le réseau.

i- Evitement du conflit entre les règles : cela est fait en équipant les métarègles avec le contexte droit nécessaire à la résolution des conflits d'attachement des syntagmes.

ii- Principe de maximisation de la couverture pour la résolution des conflits : ce principe consiste à préférer les analyses qui couvrent plus de mots. L'implantation de ce principe a été faite en donnant plus de priorité aux règles incluantes qu'aux règles incluses[40].

b. Le temps de calcul avec un RTR est cubique au pire des cas. Cependant, selon la grammaire utilisée, ce pire des cas peut ne pas être observé (En général si la grammaire ne contient pas de règles d'auto-enchâssement, le temps de calcul est linéaire par rapport à la longueur de l'énoncé). Ainsi, nous allons effectuer une analyse des temps de calcul de nos deux implantions avec les RTRs (dans le système Corrector et dans le système Oasis) afin de savoir la performance réelle de l'algorithme et la fréquence avec laquelle les pires des cas sont observés.

2.1.1.67 Résolution de problèmes particuliers

1.2.2.1.9 Modélisation de la zone d'édition

Comme nous avons vu, la zone d'édition joue un rôle particulier dans le traitement. En ce qui concerne les mots neutres (qui ne font pas partie de l'extragrammaticalité) qui apparaissent dans la phase d'édition, nous avons remarqué que ces mots jouent un rôle dans le traitement selon deux considérations :

1. **Le nombre :** le problème ici est que plus le nombre des mots neutres est élevé, plus on risque d'avoir des problèmes de surgénération. Pour éviter ce problème, nous avons décidé de ne pas accepter les patrons dont le nombre de mots neutres dépassent deux.
2. **Le sens :** selon notre observation du corpus, le sens des mots neutres joue, lui aussi, un rôle crucial dans la reconnaissance des patrons. L'exemple le plus représentatif est celui des cas d'énumération (*Two engines* ***and*** *two boxcars*) qu'on traite (incorrectement) avec des patrons comme (M1R1 X M1R1). Pour résoudre ce problème nous avons décidé d'intégrer des informations sémantiques au sein de certains patrons, qui contiennent des mots neutres, sous forme de segments conceptuels. Par exemple, le patron précité sera contrôlé par le patron (M1M2 Concept_énumération M1M2). Cette modification permet au système de reconnaître et éviter (de façon très simple) les *fausses* extragrammaticalités et de la même façon elle permet de reconnaître et de corriger des patrons qui contiennent des tournures comme *'let me see'* (dont la

[40] Une règle *X* est dite incluse dans une règle *Y* si et seulemnt si tous les symboles dans la partie droite de *X* sont inclus dans la partie droite de *Y*. Par exemple la règle $A \rightarrow B$ est incluse dans la règle $C \rightarrow BD$. Notons que *l'inclusion* est la traduction que nous avons proposée du terme *subsumption*.

longueur dépasse 2 mots). L'intégration des grammaires sémantiques est accessible puisqu'elles n'exigent que des informations de bas niveau qui portent essentiellement sur la topologie des mots.

1.2.2.1.10 Traitement des extragrammaticalités imbriquées

Comme nous avons vu dans notre étude théorique, l'imbrication est un phénomène qui implique deux extragrammaticalité partageant au moins un élément en commun. Différentes combinaisons des extragrammaticalités sont possibles. Bien que ces différentes combinaisons n'ont pas un intérêt particulier pour le modèle théorique elles ont cependant un effet direct sur le choix de la méthode pour les traiter. Ainsi, nous distinguons entre trois formes d'imbrications qui nécessitent différentes techniques de traitement :

1. **Imbrication de faux-départs avec des Els :** Nous avons vu que l'imbrication d'un amalgame au sein d'une extragrammaticalité supralexicale peut empêcher le système de la reconnaître comme dans : *I'll I will* où l'amalgame empêche l'application du patron M1M2 M1M2. Nous avons vu que ce problème est résolu par l'application de règles de traitement simples qui convertissent la forme d'amalgame en ses composantes. Par ailleurs, l'hésitation peut être une source de problème. En effet, il n'est pas rare d'observer qu'une hésitation vient se glisser au sein d'une composante syntaxique et par conséquent empêchent les règles de la l'analyser. Cela conduit à l'échec du système à reconnaître certains faux-départs dont la détection nécessite l'analyse syntaxique de leurs frontières.

   Could you give me wait I need **uh** three boxcars (102)

   Dans l'exemple précédent, le système doit analyser correctement le segment *I need uh three boxcars* afin de détecter correctement la frontière droite du faux départ (qui par définition doit être une construction bien formée) mais l'existence de l'hésitation peut l'empêcher de le faire.

   Pour résoudre ce problème, la méthode la plus simple consiste à filtrer *a priori* toutes les hésitations. Malgré son adaptation aux systèmes généralistes d'analyse comme celui décrit par (Zechner et Waibel, 1998), cette approche ne correspond pas à nos besoins. En effet, les hésitations constituent un indice important dans la détection des faux-départs et leur filtrage conduit à la perte de cette ressource. Une autre solution pour le traitement de ces phénomènes consiste à intégrer ce des modèles d'hésitations au sein même des règles syntaxiques de la grammaire. Ainsi, nous avons adopté des méta-règles pour traiter les hésitations. Les règles que nous avons utilisées sont assez proches de celles de (McKelvie, 1998) que nous avons présenté dans la troisième partie de cette thèse. Les propriétés principales des méta-règles que nous avons utilisées peuvent être résumées dans les points suivants :

   i- Les méta-règles ont été utilisées pour les constituants de base, c'est-à-dire aux règles dont la partie gauche correspond à une catégorie morpho-syntaxique : *pronpers*, *vpres*, etc.

ii- Contrairement aux règles de McKelvie qui portent sur plusieurs phénomènes comme les hésitations, certains marqueurs discursifs, etc., nos règles portent seulement sur les hésitations étant donné que le reste est traité dans le cadre des règles des faux-départ (la plupart sont représentés par la règle de la zone d'édition).

iii- Afin de prendre en considération les hésitations qui précédent un constituant ou le suivent (au début et à la fin de l'énoncé) les règles sont dotées de deux variables correspondant à des hésitations.

Ainsi, les règles utilisées sont du schéma suivant :

cat → hés* cat hés*

Dans cette règle, *cat* correspond à n'importe catégorie morpho-syntaxique et *hés* à n'importe forme d'hésitation. L'étoile signifie que le signe de l'hésitation est facultatif.

2. **Imbrication des répétitions et des autocorrections :** dans ce cas, nous avons une répétition ou une autocorrection qui est imbriquées au sein d'une autre répétition ou autocorrection. Le traitement de ces phénomènes étant fait avec des patrons cela risque de créer un conflit entre les patrons qui peuvent être appliqués à l'énoncé. Afin d'éviter ces conflits, nous avons adopté une stratégie à double niveau :

i- Le premier niveau consiste en l'application de micro-patrons (comme M1EM1, M1M1, M1-M1, M1M2 M1M2, M1M2 E M1M2) qui traitent les phénomènes simples.

ii- Le deuxième niveau consiste en l'application de tous les patrons.

Pour bien éclairer notre stratégie, nous allons examiner l'application des patrons à l'exemple d'imbrication présenté dans le premier chapitre de la troisième partie de cette thèse.

(…) <sil> do I <sil> I need two <sil> do I need two <sil> engines for the (...)

M1 M2 E M22 M3 M4 E M12 M23 M32 M42

Figure 69. Un exemple d'extragrammaticalités imbriquées

Dans le cas d'une analyse traditionnelle, le patron *M1 E M1* détecte et corrige la répétition de *I* et puis de la même façon à l'aide du patron *M1 M2 E X X M1 M2* il détecte et corrige la répétition de *need two* uniquement (le parcours de l'automate étant de gauche à droite) ce qui donne comme résultat final : *(...) <sil>* ***do*** *do i need two <sil> engines for the...*

Par contre, avec un parcours en double passe nous pouvons corriger l'énoncé correctement selon les deux étapes suivantes :

- Le micro-patron *M1 E M1* fait tout d'abord le traitement local de la répétition de *I*.
- Le patron *M1 M2 M3 M4 E M1 M2 M3 M4* traite la répétition de *do I need two*.

3. **Imbrication d'un faux-départ avec une répétition ou une autocorrection :** dans ce genre de cas, nous avons un faux-départ qui partage une partie avec une répétition ou une autocorrection. Prenons l'exemple suivant :

(…) so it must <sil> **so from <sil><brth> so from** midnight to nine a.m. (…)
(utt79, d93-11.2) (103)

Nous remarquons que l'effet de ce phénomène est similaire à l'imbrication de répétitions ou d'autocorrections que nous avons vues dans le point précédent. En effet, la reconnaissance de la partie partagée entre les deux cas (dans l'exemple précédent il s'agit de : *so from*) comme étant une partie du premier cas empêche celle de l'autre cas (en l'occurrence la répétition de *so from*). Ainsi, nous avons décidé de faire le traitement des faux-départs à partir de la deuxième passe des patrons pour bénéficier des délimitations locales de la première passe.

1.2.3 Discussion de l'architecture de Corrector

La conception de l'architecture de Corrector a été faite sur la base de différentes considérations dont les principales sont :

1. **Considérations théoriques :** comme nous avons vu dans notre analyse du *Trains Corpus* les extragrammaticalités bien qu'elles soient indépendantes des connaissances grammaticales (qui modélisent la compétence linguistique) ont une relation étroite avec ceux-ci. Ainsi, nous avons adopté une architecture qui à la fois distingue nettement les connaissances grammaticales des modèles des extragrammaticalités tout en permettant à ces sources d'information de collaborer étroitement pour traiter les extragrammaticalités. Par exemple, les informations lexicales font l'objet d'un block indépendant mais dont la sortie fait la base du traitement par patrons qui utilisent l'information morphologique dans le traitement. De même, le module d'analyse par méta-règles est indépendant du module d'analyse syntaxique partielle tout en ayant une relation privilégiée avec lui.
2. **Considérations logicielles :** comme nous vu, l'aspect principal de l'architecture d'un point de vue logiciel est l'existence d'une unité centrale (le gestionnaire de système ou le hub) autour de laquelle communiquent les différents modules. L'utilisation d'une telle architecture a plusieurs avantages d'un point de vue logiciel :
   i- **Hétérogénéité des sources d'informations :** comme nous avons vu, le système Corrector comprend sept modules répartis sur trois blocks qui couvrent des sources d'informations assez hétérogènes : lexique, patrons, méta-règles et règles syntaxiques. Ainsi, l'utilisation d'un gestionnaire de système qui est indépendant de ces sources d'informations permet d'intégrer ces différentes sources d'informations au sein du gestionnaire du système qui est indépendant de ces sources.

ii- **Portabilité :** la modularité de l'approche rend possible la réutilisation de certains modules (y compris le gestionnaire du système) dans différentes applications.

iii- **Souplesse :** la souplesse est une propriété importante dans tout logiciel quel que soit son domaine ou objectif. Dans le cas de Corrector cette propriété a influencé un bon nombre de choix (comme la localisation du système en prétraitement par rapport à un module d'analyse grammaticale). Ainsi, l'adoption d'une architecture à base de Hub rend l'intégration de Corrector au sein d'un système plus large une tâche relativement facile. En effet, tout ce dont nous avons besoin pour ce faire, est de lier le gestionnaire de Corrector au nouveau système.

1.3 Implantation du système

La partie majeure de notre système est écrite en PROLOG. Le système est composé de 7 fichiers qui correspondent à un ou plusieurs modules selon les besoins de l'implantation.

Fichier	Langage	Auteur	N.B. Lignes
Main cor double	PROLOG	M.Z.K	390
Script tagging	Perl	J. Rouillard	653
Prétraitement	PROLOG	M.Z.K.	283
Post-tag	PROLOG	M.Z.K.	314
Première passe	PROLOG	M.Z.K.	1122
Deuxième passe	PROLOG	M.Z.K.	6330
Tree drawer	PROLOG	M.Z.K.	534
Code total			8953
Total codé par nous	PROLOG	M.Z.K.	8583

Figure 70. Présentation générale du code

1.4 Exemples de traitement

Nous allons donner deux exemples de traitement de cas contenant différents types d'extragrammaticalités :

1.4.1 Premier exemple

Five a.m. okay is it faster for those for that engine to drop off those two those two boxcars travel back to Dansville than um to have engine three. **(104)**

- **Le prétraitement :** le module de prétraitement est destiné à la normalisation des amalgames ainsi que les mots oraux. Comme dans cet énoncé il n'existe pas des mots de ce genre, alors ce module rend l'énoncé tel qu'il est, sans effectuer de normalisations.
- **L'analyse morphologique :** la sortie du tagger de Xerox est la suivante :

five	+CARD
a.m.	+ADV
Ok	+ADV
Is	+VBPRES
it	+PRONPERS
faster	+ADVCMP
for	+PREP
those	+PRON
for	+PREP
that	+DET
engine	+NOUN
to	+INFTO
drop	+VINF
off	+PREP
those	+DET
two	+CARD
those	+DET
two	+CARD
boxcars	+NOUN
travel	+NOUN
back	+ADV
to	+PREP
Corning	**+PARTPRES**
than	+COTHAN
um	**+guessed+ADJ**
to	+INFTO
have	+VHINF
engine	+NOUN
three	+CARD

- **Post-tagging :** l'énoncé taggé constitue l'entrée du module suivant de post-tagging. Ce module normalise le format de la sortie (par exemple en convertissant les majuscules en minuscules, ..). De même, il corrige l'erreur de tagging du mot *Corning* (qui est dans notre contexte un nom propre pas un participe) ainsi que la catégorie associée à l'hésitation *um* considérée par le tagger comme un mot inconnu.
- **Reconnaissance locale de patrons :** la première passe traite les deux extragrammaticalités locales et fournit en sortie : Five a.m. okay is it faster for that engine to drop off those two boxcars travel back to Dansville than um to have engine three. L'autocorrection *for that for those* est traité avec le patron M1R1 M1R1' (avec R1 et R1' deux mots qui ont la même catégorie) et le répétition *those two those two* est corrigé avec le patron M1M2 M1M2. La sortie de ce module, ne contenant pas d'extragrammaticalités de niveau supérieur, le système produit en sortie l'énoncé en signalant la répétition et l'hésitation *um.*

1.4.2 Deuxième exemple

Because I have to mm maybe maybe I'll try taking um taking taking one boxcar that would be sufficient (105)

- **Le prétraitement :** ce module détecte et normalise l'amalgame *I'll* en la remplaçant par sa forme standard : *I will.*
- **L'analyse morphologique :** le tagger de Xerox fournit l'analyse suivante de l'énoncé prétraité :

because	+COSUB
I	+PRONPERS
have	+VHPRES
to	+INFTO
mm	**+MEAS**
maybe	+ADV
maybe	+ADV
I	+PRONPERS
will	+VAUX
try	+VINF
taking	+PARTPRES
um	**+guessed+ADJ**
taking	+PARTPRES

taking	+NOUNING
one	+CARDONE
boxcar	+NOUN
that	+PRON
would	+VAUX
be	+VBINF
sufficient	+ADJ

- **Post-tagging :** le seul traitement effectué par ce module est le remplacement de l'étiquette associée à l'hésitation *mm* et *um* par l'étiquette hésitation.
- **Reconnaissance locale de patrons :** ce module traite la répétition du mot *maybe* et *taking* séparément. Ceci est fait respectivement à l'aide des patrons M1M1et M1 Ed M1. Ainsi, ce module fournit la sortie suivante : Because I have to mm maybe I'll try taking taking one boxcar would be sufficient.
- **Méta-règles :** essaie d'abord les différentes règles de détection des faux départs et d'incomplétude. Il détecte et délimite le faux départ avec l'une de ces règles :

  faux_dep_segment_vpres_infto → frontière_début chunk_inc_segment_vpres_infto édition phrase_déc_inter.

  Cette règle signifie que si un segment verbal qui se termine par *to* et précédé d'une marque de début (s'il est au début de l'énoncé ou s'il est précédé d'un marquer comme l'hésitation) et suivi par une zone d'édition (hésitation ou n'importe quel autre marquer) et puis suivi par une phrase affirmative ou interrogative alors ce segment est incomplet et le cas est jugé un faux départ. Après l'examen du reste des règles et des patrons de la deuxième passe le système fournit comme sortie l'énoncé dont le faux départ est marqué (avec une délimitation des différentes zones) ainsi que l'hésitation *uh*.

1.5 Evaluation et résultats

1.5.1 Evaluation du temps de calcul de l'algorithme utilisé

Pour évaluer le temps de calcul de notre algorithme, nous avons choisi un corpus de 601 énoncés que nous avons extraits de différents dialogues.

Le graphe suivant montre la fréquence des énoncés dans notre corpus comparée à leurs longueurs :

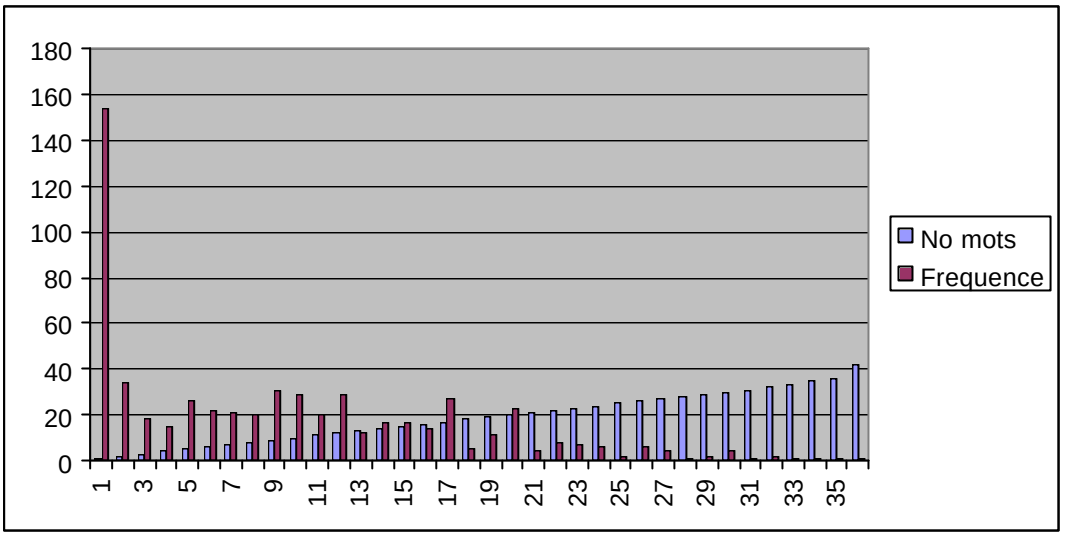


Figure 71. Les fréquences des énoncés utilisés dans le test comparées à leurs longueurs

Comme nous pouvons le remarquer dans le graphe précédent, les énoncés avec un seul mot ont une occurrence assez considérable comparés aux autres. Il s'agit globalement de connecteurs discursifs comme *oui, non, ok*, etc. Nous remarquons aussi, qu'à partir de 20 mots, la fréquence des énoncés commence à baisser.

Les expériences ont été faites sur un PC Pentium III/500 Mega hertz et 196 KB de RAM. Les temps de calculs considérés portent uniquement sur les modules de post-tagging parce que le tagger peut être vu comme un module externe à l'étiquetage proprement dite des extragrammaticalités d'une part et d'autre part cela permet d'éviter les biais qui peuvent résulter du lien entre le tagger et le site de Xerox à travers Internet.

Afin de donner une idée sur le comportement réel du système nous avons décidé de montrer les performances du système selon deux critères différents :

- La moyenne du temps de calcul.
- Les pires des cas observés pour chaque longueur.

2.1.1.68 La moyenne des temps de calcul

Selon nos calculs, la moyenne générale du temps de calcul par énoncé est de 7, 61 secondes. Cette moyenne générale ne donnant qu'un indice général du comportement du système, nous avons décidé de calculer les moyennes de temps de calcul pour chaque longueur d'énoncé. Les résultats de notre tableau sont présentés dans le graphe suivant :

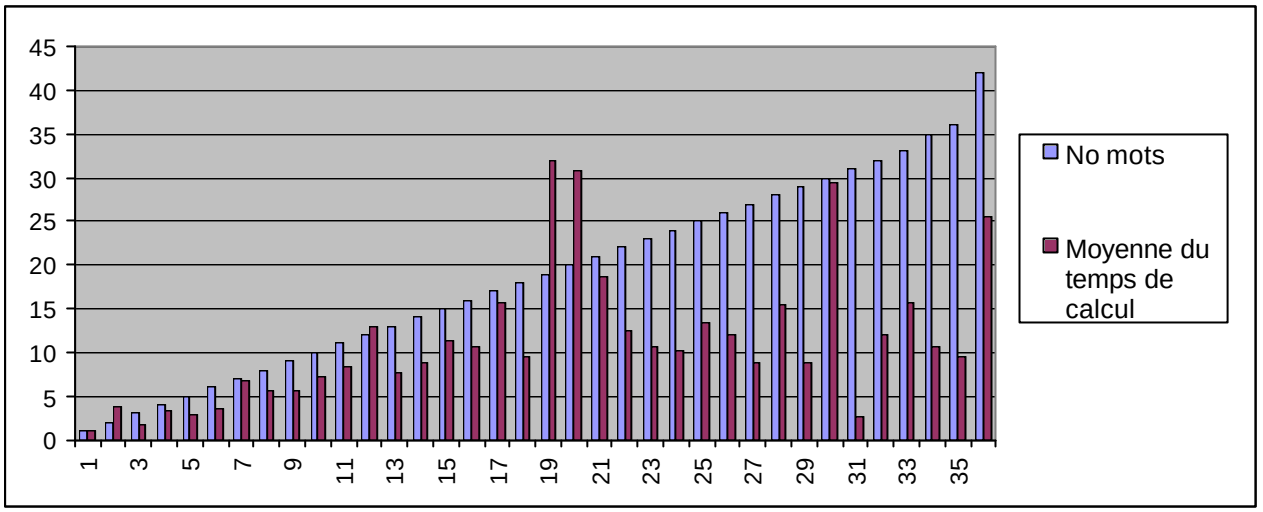


Figure 72. Les moyennes des temps de calcul observés

Comme nous pouvons le remarquer dans la figure précédente, les temps de calculs augmentent graduellement jusqu'à atteindre leur sommet aux environ de 20 mots et puis ils baissent globalement sauf dans deux cas. La raison pour laquelle l'augmentation du temps de calcul n'est pas systématique est une combinaison des facteurs longueurs et fréquence. En effet, les énoncés les plus courts sont très fréquents mais vue leur longueurs ils ne permettent pas d'observer des augmentations significatives dans les temps de calcul. Par contre, les énoncés aux environs de 20 mots sont à la fois assez fréquents et suffisamment longs pour que les pires des cas de la complexité de l'algorithme soient observés dans leur cadre.

2.1.1.69 Les pires des temps de calcul observés

La considération des pires des temps observés permet de donner une idée sur le comportement de l'algorithme dans conditions extrêmes observées dans notre corpus. Le graphe correspondant au pire des cas observés dans notre corpus d'évaluation est présenté dans la figure suivante :

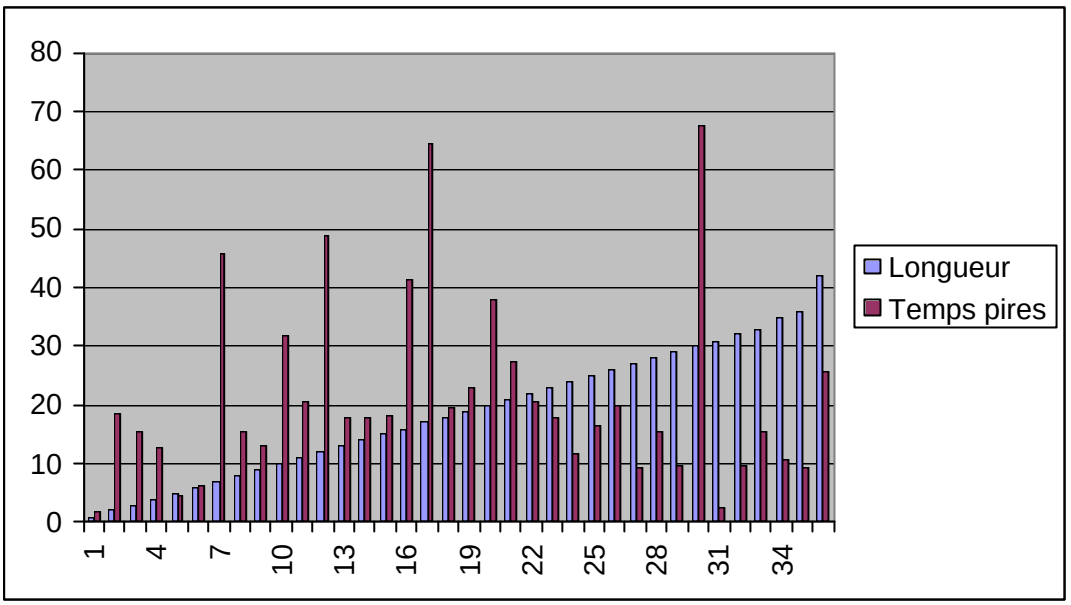


Figure 73. Les temps de calcul obtenus sur les pires des cas observés

Le graphe des temps obtenus sur les pires des cas observés confirme notre constatation avec les moyennes des temps. Nous remarquons que l'augmentation du temps de calcul ne dépend pas de manière systématique de la longueur de l'énoncé analysé. Par ailleurs, ce graphe montre que l'augmentation des moyennes dans le graphe précédent dépend d'augmentations ponctuelles dans des cas particuliers plutôt que d'être le résultat d'une tendance générale.

1.5.2 Evaluation du traitement des extragrammaticalités

La première étape de notre évaluation a consisté en la construction du corpus de test. Il s'agit de 305 énoncés non utilisés pour l'écriture de la grammaire. Parmi ces énoncés, 255 contiennent 309 extragrammaticalités supralexicales. Les cinquante autres énoncés ne contenant pas d'extragrammaticalités supralexicales ont été utilisés pour tester la surgénération du système. Le pourcentage des phénomènes dans les énoncés correspond approximativement à 30% du nombre total des phénomènes observés dans le corpus. Ainsi, nous avons 76 cas d'autocorrections, 91 cas de faux départs, 52 cas d'incomplétudes et 90 cas de répétitions. Outres les extragrammaticalités supralexicales, notre corpus de teste contient 272 cas d'extragrammaticalités lexicales.

Les résultats de notre test sont présentés dans la figure suivante :

Phénomène			%
Extragrammaticalités lexicales	Détection	Rappel	98,89
		Précision	98,17
	Délimitation	Rappel	98,89
		Précision	98,17
Répétitions	Détection	Rappel	96,20
		Précision	98,75
	Délimitation	Rappel	91,13
		Précision	90
Autocorrections	Détection	Rappel	77,55
		Précision	92,68
	Délimitation	Rappel	71,42
		Précision	85,63
Faux-départs	Détection	Rappel	68,08
		Précision	78,04
	Délimitation	Rappel	53,19
		Précision	60,97
Incomplétude	Détection	Rappel	85,71
		Précision	80
	Délimitation	Rappel	71,42
		Précision	66,66
Total extragrammaticalités supra lexicales[41]	Détection	Rappel	81,56
		Précision	88
	Délimitation	Rappel	71,79
		Précision	76,44
Total	Détection	Rappel	89,67
		Précision	92,76
	Délimitation	Rappel	84,47
		Précision	86,61

Figure 74. Résultats du système Corrector sur le corpus test

[41] Pour calculer les pourcentages des extragrammaticalités supralexicales ainsi que les pourcentages totaux, nous avons additionné les pourcentages des phénomènes multipliés chacun par son nombre d'occurrences dans notre corpus de test et puis nous avons divisé le tout sur le nombre total des occurrences.

2.1.1.70 Analyse des résultats

Voici une analyse des résultats du système Corrector organisée selon les principaux phénomènes qu'il couvre :

- **Les extragrammaticalités lexicales**

Comme nous remarquons dans le tableau précédent, les taux de reconnaissance des extragrammaticalités lexicales sont assez élevés. Cela montre que la couverture de notre grammaire était assez bonne d'un part et d'autre part, que notre approche était bien adaptée. Nous remarquons aussi que les taux de reconnaissance et de délimitation sont identiques puisqu'il n'existe pas un problème lié à l'étendue d'une extragrammaticalié lexicale.

- **Les répétitions**

Les répétitions, comme nous le remarquons dans le tableau, ont été traitées avec des taux assez élevés tant pour la détection que pour la délimitation. Nous pouvons remarquer aussi que la précision est bonne ce qui montre que notre approche pour la réduction des surgénérations a donné ses fruits. Environ 70% des erreurs de notre système sont causées par des zones d'édition très complexes qui impliquent non seulement des éditeurs mais aussi des mots normaux. Le reste est principalement dû à des problèmes de sous-génération des patrons. Par ailleurs, quatre des cinq cas de répétitions imbriquées que nous avons observés dans notre corpus de test ont été traités correctement. Le cas non-traité contient une erreur d'analyse morphologique qui est la raison de l'échec.

- **Les autocorrections**

Comme nous pouvons le remarquer dans le tableau précédent, le rappel des autocorrections est moins élevé que celui des répétitions alors que les taux de précisions sont plus proches. Les principales sources d'erreurs sont les suivantes :

- Les erreurs d'analyse morphologique (40% des erreurs).
- Sous-génération des patrons constituent 40% des erreurs. Le tiers de ces erreurs, c'est-à-dire, 10% du total est dû à des cas très compliqués. Parmi les cas difficiles nous pouvons citer les autocorrections impliquant non pas le remplacement d'un mot par un autre mais plutôt le remplacement d'un mot par une unité syntaxique ou l'inverse comme dans l'exemple suivant : *I'm gonna take I'm taking*.
- Problèmes liés à la zone d'édition (20% des erreurs). Ces erreurs sont dues à des formes de la zone d'édition non modélisables syntaxiquement ou sémantiquement comme l'insertion d'un verbe ou nom au sein de la zone d'édition.

Nous remarquons que la précision est assez élevée. Cela montre, encore une fois, que nos tentatives de réductions de surgénération ont donné leurs fruits. Parmi les dix cas d'extragrammaticalités imbriquées deux cas seulement n'ont pas été correctement traités.

- **Les faux-départs**

Les taux de rappel et de précision des faux-départs sont moins élevés que ceux dans les deux cas précédents. Ceci est dû à la fois à la complexité de ces phénomènes et à la richesse des informations qui ont été utilisées pour les traiter (par exemple on est plus dépendant de l'analyseur morphologique que dans les autocorrections). Les raisons principales des erreurs de traitement de ces phénomènes sont les suivantes :

i. La sous-génération (55%).

ii. Cas très compliqués (25%). Ces cas sont principalement dus à des verbes qui peuvent être tantôt transitifs et tantôt intransitifs ainsi qu'à l'imbrication de plusieurs extragrammaticalités (plus de deux cas).

iii. Des erreurs d'analyse morphologique (20%).

- **Les incomplétudes**

Nous remarquons que les taux de détection et délimitation des incomplétudes sont plus élevés que ceux des faux départs. Cela est motivé par le fait que la frontière droite est par définition délimitée dans les incomplétudes, ce qui facilite à la fois la détection et la délimitation de ces phénomènes. 62,5% des erreurs observées sont dues à des problèmes de sous-génération de notre corpus alors que dans 37,5% les erreurs d'analyse morphologique était la source de l'erreur d'analyse.

2.1.1.71 Comparaison avec le système de Heeman

Dans ce qui suit, nous allons comparer nos résultats à ceux de Peter Heeman (Heeman, 1998). Ce choix est motivé par les trois raisons suivantes :

- Ce travail est basé aussi sur le *Trains Corpus* que nous avons utilisé pour notre système.
- Nous couvrons pratiquement les mêmes phénomènes à l'exception des incomplétudes et des extragrammaticalités lexicales.
- A notre connaissance, les résultats obtenus par Heeman sont les meilleurs dans la littérature pour les tâches de détection et délimitation combinées.

Malgré tous ces facteurs rapprochants, il n'est cependant pas inutile de rappeler que cette comparaison est approximative dans la mesure où nos corpus de test ne sont pas identiques d'une part et d'autre part, parce que les conditions de test en général et les définitions des phénomènes ne sont pas les mêmes.

Comme nous avons vu dans le deuxième chapitre de la deuxième partie, la typologie de Heeman distingue trois types de phénomènes :

1. **Les discontinuités (*abridged repairs*) :** ce terme couvre les hésitations et les mots incomplets. Les résultats obtenus par Heeman ne sont pas comparables aux nôtres parce que d'une part, nous ne considérons pas les mots incomplets dans nos tests étant donné qu'ils ne peuvent pas

être reproduits par les systèmes de reconnaissance et d'autre part, Heeman ne considère pas les amalgames comme une forme d'extragrammaticalité lexicale.

2. **Les réparations (*modification repairs)* :** le terme réparations couvre à la fois les répétitions et les autocorrections. Si nous calculons la moyenne de nos résultats sur ces deux phénomènes nous obtenons 86,87% de rappel pour la détection et 95,71% de précision pour la détection. Comparés aux résultats obtenus par Heeman (rappel 80,87% et 83,37% précision) nous remarquons que notre système présente un avantage d'environ 6% au niveau du rappel et environ 12% pour la précision. Quant à la délimitation, la moyenne obtenue est 81,27% pour le rappel et 87,81% pour la précision. Si nous comparons ces résultats à ceux obtenus par Heeman (77,95% pour le rappel et 80,36% pour la précision), nous remarquons que nous avons une amélioration d'environ 3% pour le rappel et 7% pour la précision. Ces améliorations peuvent être justifiées par les deux points suivants :
   - Avantages Sur le plan du rappel (la couverture) : l'augmentation de la couverture est, en partie, due à l'augmentation des patrons d'environ 40% que nous avons effectuée en générant et ajoutant analogiquement de nouveaux patrons.
   - Avantages en ce qui concerne la précision : l'augmentation de la précision est due aux différents aspects de notre approche visant à réduire les surgénérations comme : les patrons de contrôle, l'ordonnance des patrons, modélisation de la zone d'édition et la double passe.
3. **Les faux départs :** en ce qui concerne les faux départs, nous avons obtenu pour la détection un rappel de 68,08% et une précision de 78,04%. Cela veut-dire que nous avons réalisé un avancement d'environ 20% pour le rappel et de 9% approximativement pour la précision par rapport aux résultats obtenus par Heeman (48,58% de rappel et 69,21% de précision). En ce qui concerne la délimitation, nous avons obtenu 53,19% de rappel et 60,97% de précision. Ainsi, nous avons obtenu une amélioration d'environ 17% pour le rappel et 9% approximativement pour la précision (Heeman a obtenu 36,21% de rappel et 51,59% de précision). Cette amélioration est justifiée à la fois par la prise en considération des propriétés syntaxiques des faux départs, en particulier la prise en considération du contexte droit qui est un facteur décisif pour la détection d'un bon nombre d'extragrammaticalités. De plus, notre approche d'analyse partielle par segment (qui a servi de base aux méta-règles des faux-départs) s'est montrée assez robuste même dans les cas d'extragrammaticalités.

En ce qui concerne l'incomplétude, nous ne pouvons pas comparer nos résultats à d'autres travaux parce que comme nous avons dit dans les chapitres précédents, à notre connaissance, ce phénomène n'a pas fait explicitement l'objet d'une étude ou d'une implantation.

1.6 Bilan du système Corrector

Nous avons présenté dans ce chapitre notre système Corrector qui est la réalisation pratique de notre modèle des extragrammaticalités. Comme nous avons vu, les points clés de ce système sont les suivants :

1. **Niveau lexical :** au niveau lexical, nous avons adopté une approche qui vise trois objectifs principaux :
   i. Réduction des erreurs de détection et délimitation d'extragrammaticalités supralexicales dues à des extragrammaticalités lexicales comme dans : *I'll uh I will*.
   ii. Minimisation des erreurs d'analyse morphologique, d'une part en choisissant un tagger adapté et d'autre part, en effectuant les prétraitements et les post-traitements qui permettent de réduire les erreurs de ce tagger.
   iii. Réduction des effets des erreurs d'analyse morphologique : sur ce plan, nous avons utilisé des techniques qui ne nécessitent pas le recours systématique à l'information morphologique comme la reconnaissance de patrons ou des techniques d'analyse superficielle qui tolèrent certaines erreurs d'analyse morphologique.
2. **Le niveau des patrons :** notre approche d'analyse symbolique nous a permis d'augmenter les patrons observés en générant de nouveaux patrons de manière analogique. Comme nous avons vu cela nous a permis d'avoir 40% de patrons de plus que nous avons observés. Nous avons adopté une approche qui réduit considérablement les conflits potentiels entre les patrons.
3. **Les règles syntaxiques :** nous avons adopté une approche fine qui prend en considération une catégorisation particulière qui permet d'exprimer à la fois des contraintes très fines (syntagme nominal composé d'un pronom personnel sn_pron_pers), ou des contraintes générales du type SN, SV, etc. Pour réduire la sous-génération, nous avons utilisé plusieurs procédures comme les règles et patrons de contrôle, l'utilisation du contexte pour contraindre le système dans la considération de certains *segments relativement extragrammaticaux* comme étant des faux-départs.
4. **Techniques diverses :** nous avons utilisé différentes techniques pour augmenter la couverture au maximum tout en réduisant la sous-génération. Parmi ces techniques nous pouvons citer : l'utilisation des grammaires sémantiques pour la modélisation de certaines formes des zones d'édition et l'adoption d'une approche à double passe pour l'étiquetage des extragrammaticalités imbriquées.

Nous avons vu que notre évaluation a confirmé généralement les avantages théoriques que nous avons présentés. En effet, nous avons obtenu des résultas meilleurs que ceux de Heeman tant pour le rappel que pour la précision. L'amélioration la plus importante était dans le traitement des faux départs. Cela montre d'une part, la pertinence de nos remarques sur les travaux de Heeman qui ont utilisé des N-

grams ainsi que sur les travaux de Core qui ont eu recours à des règles syntaxiques qui ne prennent pas en considération suffisamment de contexte pour contraindre les segments jugés extragrammaticaux. Plus généralement, nos résultats ont montré que les informations syntaxiques constituent un indice important non seulement pour la délimitation des faux départs mais aussi pour leurs détections.

Le bilan des raisons d'erreurs de notre système peut être résumé dans les points suivants :

- La sous-génération est la raison principale des erreurs de notre système.
- Les erreurs d'analyse morphologique constituent une source importante d'erreurs. Cependant, comparé au niveau général des résultats, nous remarquons que notre approche a permis de limiter l'effet de ces erreurs à niveau acceptable.

2 Chapitre III.2 : Les systèmes Safir et Oasis pour l'analyse du langage oral dans le contexte de dialogues orientés par la tâche

Nous allons présenter dans ce chapitre les implantations des formalismes S-TSG et Sm-TAG. Il s'agit respectivement des systèmes Safir et Oasis.

2.1 Les premiers pas : le système SAFIR

Le système SAFIR a été le premier pas dans notre travail sur l'analyse robuste du langage oral (il a été réalisé avant les systèmes Corrector et Oasis). Bien qu'il s'agit plus d'un prototype que d'un travail complètement finalisé, nous avons jugé bon de le présenter dans ce document afin de donner au lecteur une idée sur la base et les motivations des choix que nous avons fait plus tard dans la nouvelle version du système baptisée Oasis à laquelle nous allons consacrer les chapitres suivants de cette partie.

2.1.1 Le corpus de réservation hôtelière

Le système SAFIR est construit sur le corpus de Réservation hôtelière qui a été collecté au sein de l'équipe GEOD du laboratoire CLIPS-IMAG. La collecte de ce corpus a été faite en suivant la méthode de la simulation dialogique (Hollard, 1997). Les dialogues obtenus portent sur des questions sur la disponibilité, le prix, les propriétés des chambres de même que les dates d'arrivée ou de départ des clients, des expressions de politesse, etc. Le corpus contient 184 dialogues qui font 166 Kb de données (31376 mots). Parmi ces dialogues, 148 dialogues ont abouti à une réservation réussie. Les autres représentaient soit une demande de réservation non satisfaite soit d'autres demandes : renseignements (sur le prix, le trajet), complément d'une réservation précédente (modification ou annulation).

En moyenne, chaque dialogue contient 7,28 énoncé de client, ce qui fait un total d'environ 1339 énoncés de client dans ce corpus.

L'avantage principal de ce corpus est son adaptation sémantique et pragmatique puisque les énoncés produits par les sujets reflètent fidèlement la tâche du dialogue : acte de dialogue de demande de réservation, informations, prix, etc. Sur le plan syntaxique, la syntaxe des énoncés produits est très proche de celle que nous avons observée dans d'autres corpus de tâches différentes. L'inconvénient principal de ce corpus est l'absence presque totale des extragrammaticalités, à l'exception de quelques hésitations et autres extragrammaticalités lexicales nous avons rarement observé des cas de répétition,

d'autocorrection de faux-départ ou d'incomplétude dans ce corpus. Un exemple d'un dialogue extrait de ce corpus est présenté dans l'annexe 1.

2.1.2 Les requis du système

Ayant affaire à des dialogues oraux finalisés, les points que nous devons prendre en considération lors du choix tant de l'architecture du système que de la nature des composantes peuvent se résumer comme suit :

- La limitation de la tâche du dialogue : le nombre du lexique nécessaire pour le traitement de la tâche du dialogue est assez limité. Ainsi, les ambiguïtés lexicales que nous pouvons avoir sont aussi très limitées.
- On a de bonnes possibilités de prédictibilité d'événements tant linguistiques que pragmatiques, étant donné que la tâche du dialogue (la réservation touristique) est relativement limitée.
- On aura affaire à un bon nombre de phénomènes d'extragrammaticalité (hésitation, incomplétude, etc.) dus à la spontanéité de la parole.
- Même si, à ce stade, nous allons travailler sur des énoncés transcrits, nous devons prendre en considération les erreurs de reconnaissance de la parole lors du choix de la stratégie.

2.1.3 Architecture du système

Nous avons choisi d'intégrer les différentes composantes du système au sein d'une architecture sérielle. La motivation principale du choix de cette architecture est sa modularité et sa simplicité. En effet, une approche modulaire permet la création de modules spécialisés pour chacune des sous-tâches de traitement et donne, par conséquent, plus de souplesse pour la substitution des différentes composantes du système si l'une d'elles s'avère moins adaptée que les autres.

Les propriétés clés de SAFIR sont les suivantes :

- L'entrée du système est les transcriptions des énoncés.
- Un module d'analyse basé sur le formalisme S-TSG.
- La sortie du système est une représentation sémantique superficielle sous forme de schéma.

Comme le montre la figure 75, l'architecture de SAFIR est composée de trois modules principaux :

2.1.1.72 Justification des choix

1. Sur le plan morphologique, nous pouvons bénéficier de la limitation du lexique en stockant toutes les formes utiles des mots pertinents pour la tâche. Cela nous évite la création d'un analyseur morphologique.
2. Pour bénéficier de la limitation de la tâche et éviter les extragrammaticalités de l'oral, nous allons procéder à une analyse partielle de l'entrée. En d'autres termes, on ne va chercher dans le message que les réalisations (des concepts) pertinentes pour la tâche. Cette phase sera assurée par un ensemble d'arbres locaux.

3. Le niveau des arbres globaux nous permet de lier les structures obtenues avec les arbres locaux.
4. Comme représentation sémantique finale, nous avons choisi le formalisme des schémas (Minsky, 1975). Le choix de ce formalisme est justifié par sa simplicité ainsi que la profondeur acceptable qu'il permet pour l'application générale visée par Safir.

Ainsi, nous avons proposé une architecture en trois modules :

- Le prétraitement.
- L'analyse linguistique.
- L'analyse sémantique (les schémas).

L'architecture générale de Safir est présentée dans la figure suivante :

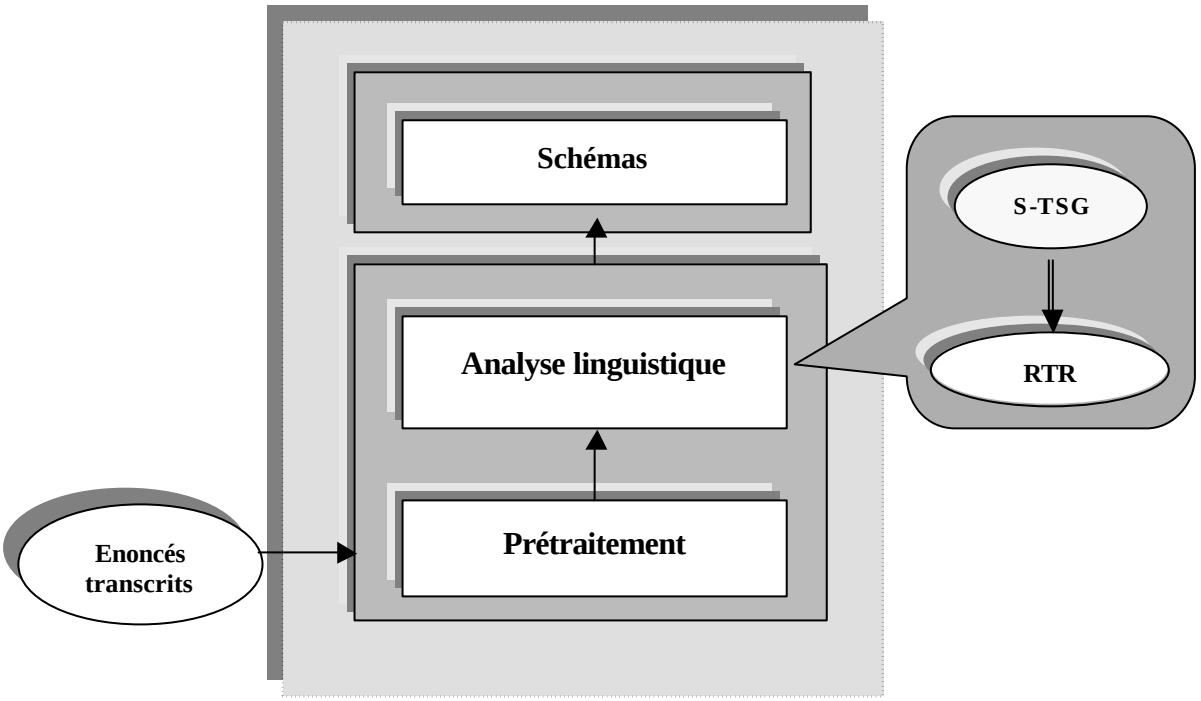


Figure 75. Architecture du système SAFIR

2.1.1.73 Le prétraitement

La fonction principale du prétraitement est de faire une série d'adaptation du format de l'entrée (comme la conversion des chiffres en mots) au format de la grammaire.

2.1.1.74 L'analyse linguistique

Ce module est basé sur une grammaire S-TSG convertie en un RTR enrichi par une stratégie sélective. Les démarches de l'écriture de la grammaire ainsi que les différentes propriétés de la méthode d'analyse seront présentées dans les paragraphes suivants.

2.1.3.1.1 L'écriture de la grammaire

Dans cette phase, notre démarche était essentiellement onomasiologique, c'est-à-dire notre objectif était de chercher toutes les réalisations possibles d'un concept ce qui nous permet de faire un filtrage

préliminaire des réalisations pertinentes pour la tâche. L'analyse du corpus a été faite selon deux étapes :

1. **Création d'une ontologie préliminaire de la tâche :** l'objectif de cette étape est de créer une base des concepts clés dans la tâche et de représenter leur relations (pour une présentation pratique du processus de construction des ontologies voir (Noy et Mcguiness, 2001)). Sur la base d'un sous-corpus de vingt dialogues, nous avons crée une version préliminaire de l'ontologie de la tâche. Au début de l'analyse, nous avons segmenté conceptuellement un ensemble de phrases pertinentes pour la tâche et appartenant à 20 dialogues.
2. **Etiquetage des données selon l'ontologie préliminaire :** nous avons appliqué l'ontologie crée au reste des dialogues. Ceci a été fait en classifiant les segments de chaque énoncé selon l'ontologie préliminaire et en l'enrichissement de cette-dernière par de nouveaux concepts jugés utiles pour la tâche et non couverts par l'ontologie préliminaire. La liste finale compte 36 concepts. La transcription des énoncés a été faite selon les symboles présentés dans le tableau suivant.

Symboles utilisés	Informations notées
→	Contexte immédiat à droite
←	Contexte immédiat à gauche
→ /	Contexte lointain à droite
/ ←	Contexte lointain à gauche
C	Contexte voisin émis par le client
H	Contexte voisin émis par l'hôtelier
[]	Segment correspondant à un arbre local

Tableau 11. Symboles utilisés pour l'annotation du corpus

Au cours de l'annotation nous avons observé que certains concepts sont très fréquents (existent presque toujours dans les dialogues) par exemple : *formule_de_demande*, *salutation_ouverture*, etc. alors que certains autres sont relativement rares comme : *annulation de la réservation*. Par ailleurs, sur le plan de la richesse on avait des concepts dont les réalisations sont très nombreuses alors que les réalisations de certains autres sont moins variées.

Finalement, sur le plan de la pertinence nous pouvons diviser nos concepts en deux parties :

1. Des concepts prototypiques, c'est-à-dire l'ensemble des concepts qui sont à la fois nécessaires et valables pour le traitement de tous les dialogues comme par exemple le concept *formule de demande*.
2. Des concepts qui sont pertinents pour la tâche et dont la réalisation dépend du client et de l'idée qu'il a, *a priori*, de l'hôtel comme par exemple le concept *emplacement de la chambre :* les

clients qui ont une idée sur l'hôtel, demandent parfois des chambres qui donnent sur une rue quelconque, sur un lac, etc.

Par ailleurs, nous avons observé plusieurs cas difficiles au cours de l'annotation. Par exemple, dans certains énoncés, nous avons eu des cas que nous pouvons qualifier d'*amalgame conceptuel* comme dans : *Vous reste-t-il des chambres*, où nous sommes pratiquement incapables de distinguer la partie qui concerne l'interrogation de celle de la disponibilité. Pour résoudre ce problème, nous avons crée une classe supplémentaire qui correspond à celles des deux arbres locaux non-amalgamés.

2.1.3.1.2 L'implantation de la grammaire

La S-TSG a été convertie en un RTR. Les principales motivations de ce choix sont présentées dans les points suivants :

- La S-TSG étant fortement équivalente à une CFG, il est formellement possible de convertir toute S-TSG en un RTR (qui est à son tour fortement équivalent à une CFG).
- Facilité d'implantation du RTR et avantages en termes de visualisation des arbres sous forme de réseaux.
- Approche descendante qui permet d'implanter les prédictions basées sur la tâche.

La S-TSG étant fortement équivalente à une CFG, la procédure de conversion revient à convertir une CFG normale en un RTR (pour plus de détails voir plus loin la conversion des arbres à substitution en RTRs dans le paragraphe du système Oasis).

Les RTRs que nous avons utilisés ont deux spécificités qui les rendent plus adaptés au traitement de l'oral. Il s'agit de l'analyse partielle et de la stratégie sélective.

1. **Analyse partielle :** l'approche d'analyse partielle consiste à permettre à des structures partielles d'être considérées comme des analyses correctes. Ainsi, selon cette approche, le système essaie tout d'abord de trouver une analyse de l'entrée avec un réseau qui correspond à un arbre global si cela s'avère impossible il accepte d'analyser l'entrée avec une série d'arbres locaux séparés. Parfois le système combine les réseaux globaux et locaux dans l'analyse du même énoncé.
2. **La stratégie sélective :** la fonction principale de cette stratégie est de localiser les zones pertinentes dans l'entrée et afin de permettre au système de les traiter. Une telle approche a plusieurs avantages comme nous avons vu dans la deuxième partie. En effet, elle permet de réduire la sous-génération de la grammaire, réduit le traitement aux seules zones pertinentes. Nous avons testé deux approches pour la localisation des zones pertinentes : la grammaire de nettoyage et l'algorithme de détection des frontières des arbres.
    - **A. Grammaire de nettoyage :** il s'agit d'un ensemble d'heuristiques que nous avons proposées pour modéliser les segments que le système est incapable d'analyser. Ces heuristiques sont classées selon l'emplacement dans le traitement des segments non analysables et se complètent entre elles par une stratégie coopérative.

i- Les heuristiques : nous avons utilisé quatre heuristiques différentes. Voici leur description détaillée :

a- **Heuristique initiale** : elle a la forme suivante :

[n'importe quel mot : d[42]] [mot_bruit : f]

Cette heuristique permet d'ignorer n'importe quel mot (qu'il soit du lexique ou pas) à condition qu'il figure au début de la chaîne et qu'il soit immédiatement suivi d'au moins un mot bruit.

b- **Heuristique intermédiaire** : cette heuristique a la forme suivante :

[mot_bruit : +d] [n'importe quel mot] [mot_bruit : f]

Sont considérés comme étant du bruit, tous les mots qui figurent entre, au moins, deux mots bruits. Cette règle permet de consommer à la fois le mot non pertinent et tous les mots bruits qui viennent avant lui et un seul mot bruit de ceux qui peuvent venir après.

c- **Heuristique finale :** cette heuristique a la forme suivante :

[mot_bruit : +d] [n'importe quel mot] [mot_bruit : +f]

Etant donné que cette règle est destinée à reconnaître les réalisations du bruit qui figurent à la fin de la chaîne, elle autorise en plus de la règle précédente la consommation de tous les mots bruits qui figurent à la fin de la chaîne.

d- **Heuristique finale bis :** le schéma général de cette heuristique est le suivant :

[mot_bruit : +d] [n'importe quel mot]

Cette heuristique permet d'ignorer tous les mots *bruit* qui peuvent figurer avant le mot lexique et le mot lui-même. Elle est particulièrement efficace pour nettoyer les mots non pertinents qui figurent à la fin de la chaîne.

ii- La stratégie coopérative : il s'agit d'un ensemble de règles qui contrôlent l'interaction des différentes heuristiques afin d'augmenter leur efficacité pour le nettoyage et réduire les conflits entre elles. Pour mettre au clair cette stratégie, nous allons présenter la position des différentes heuristiques au sein d'une règle globale composée de deux états (d'un RTR) :

[42] Les symboles ajoutés à la fin des segments indiquent la localisation de ces segments dans la chaîne d'entrée. Ils ont la signification suivante : *d* pour début et *f* pour final. Par ailleurs, nous avons utilisé une étoile (*) pour marquer un élément facultatif et le symbole (+) pour marquer les éléments qui peuvent se répéter.

[heuristique initial *+ : d] ←→ [heuristique intermédiaire * : +d] [réseau1 : d] [heuristique intermédiaire* : +] [réseau 2 : f] [heuristique intermédiaire * +] [heuristique finale 1* f] [heuristique finale2* : f].

La première remarque qu'on peut faire à propos de cette règle est que tous les réseaux de bruit sont facultatifs. Cela qui veut dire que ces réseaux n'imposent pas de contraintes qui peuvent alourdir les règles ou les empêcher de reconnaître un élément qu'elles pourraient reconnaître si elles n'étaient pas équipées d'une stratégie sélective.

Pour représenter l'aspect fonctionnel de notre stratégie coopérative, nous allons la diviser en trois blocs :

a- Partie initiale : comme nous l'avons dit, cette partie est conçue pour traiter les débuts de chaînes. Ainsi, les heuristiques de ce bloc peuvent traiter, à côté de la chaîne de bruit pur (qui peut être traitée soit par les règles locales soit par les règles globales), des cas assez variés. En voici une présentation générale [43] :

- Des chaînes du type : BL BL […]

  Ces chaînes peuvent être traitées par la première heuristique qui, grâce à sa capacité de répétition, consomme tout d'abord les deux premiers mots et ensuite les deux qui restent.

- BL [..] L(…)BL

  Une telle chaîne peut être traitée tout d'abord par la règle initiale et ensuite par la règle intermédiaire.

- Des chaînes comme BL […] L(…)BL […] L

  Cette chaîne peut être traitée par trois règles : la règle initiale consomme les chaînes du type BL, la règle intermédiaire consomme les chaînes du type LBL, et finalement la chaîne de bruit pur sera traitée par la stratégie de saut placée à la tête des grammaires locales. Ici, on remarque la raison pour laquelle le deuxième état bruit du réseau intermédiaire ne peut pas se répéter. En fait, cela a l'avantage de donner plus de chance à un autre réseau intermédiaire de s'activer (puisque ce deuxième a besoin d'au moins un mot bruit au début pour pouvoir s'activer).

- BL […] L(...)BL BL

  h. ini h. int h. ini

[43] Les mots bruits seront symbolisés par ***B*** et les mots du lexique seront symbolisés par ***L***. La répétition de la même chaîne est représentée par [..] ou même caractère par (…). Les espaces entre les sous chaînes séparent les segments qui sont traités par la même règle en une seule itération.

A la différence de la chaîne précédente, la chaîne BL, à la fin, ne peut pas être traitée par la règle intermédiaire, ce qui implique, un retour arrière vers la règle initiale.

b- **Partie intermédiaire :** en général, cette partie est moins exposée au bruit que les deux autres puisqu'elle figure entre la réalisation de deux arbres locaux (qui sont censées être liées étroitement) et elle est, en outre, plus limitée par le fait qu'on ne peut pas consommer les mots *bruit* qui figurent à la fin ou au début de la chaîne.

c- **Partie finale :** cette partie a été conçue pour traiter, à côté des chaînes (LBL) qu'on vient de voir, des chaînes de deux types :

- Des chaînes qui se terminent avec plusieurs mots bruits. Ces chaînes posent des problèmes à la règle intermédiaire dont le dernier état (bruit) ne peut pas se répéter pour la raison qu'on vient d'expliquer. Pour résoudre ce problème, nous avons proposé *l'heuristique finale1* qui consomme le lexème bruit et tous les mots bruit qui viennent avant et après.
- Des chaînes qui se terminent par un mot du lexique. Ces chaînes, qui ne peuvent pas être traitées par la règle intermédiaire, sont traitées par la règle finale bis.

Enfin, pour traiter les cas où deux mots du lexique se succèdent dans une position non-pertinente, nous proposons le recours à des modèles des réalisations les plus fréquentes des bi-mots non-pertinents, et leur intégration au sein de notre stratégie de nettoyage.

B. L'algorithme de détection des frontières des arbres : cet algorithme est basé sur deux sources d'informations :

i- L'aspect descendant de l'analyseur.

ii- La frontière lexicale FL des réseaux locaux. Par FL nous entendons, le premier élément lexical dans le réseau après la satisfaction des transitions. Par exemple, lorsque le système prédit un réseau qui correspond à un arbre local, il prend la liste de tous les FL possibles de cet arbre comme référence et compare tous les mots de l'entrée aux éléments de cette liste. Si le mot ne fait pas partie des FLs du réseau prédit, il est immédiatement ignoré et le processus est renouvelé avec le reste des mots jusqu'à ce qu'on trouve un item lexical dans l'entrée qui fait partie des FLs et à ce moment là on commence l'analyse. Sinon, on continue jusqu'à ce qu'on épuise tous les éléments lexicaux de l'entrée. Trois heuristiques légèrement différentes sont

utilisées pour la sélection des zones pertinentes d'un message. Voici un exemple d'une heuristique simplifiée :

Pour chaque séquence d'entrée S et un réseau prédit R_1 ;
Soit FL_{R1} la liste des mots;
Comparer le premier mot dans l'entrée w_1 aux unités lexicales de FL_{R1} ;
Si w_1 fait partie de de FL_{R1} ;
Alors commencer l'analyse ;
Sinon, ignore-le;
Répéter le processus jusqu'à trouver un mot w_x qui fait partie de FL_{R1} ;
Si tous les mots de S ne font pas partie de FL_{R1}
Alors, recommencer le processus avec le deuxième arbre R_2 prédit par le système.

Figure 76. Une version simplifiée de l'heuristique sélective

Selon nos tests informels, cette approche s'est avérée trop inefficace d'un point de vue calcul et donc a été abandonné.

2.1.4 Implantation du système

Safir a été implanté en PROLOG. Le choix de PROLOG est motivé par l'adaptation de ce langage au traitement symbolique ainsi que la rapidité du développement qu'il permet. La longueur totale du code est de 1939 lignes.

2.1.5 Evaluation et résultats

Pour évaluer le système, nous avons utilisé des énoncés 327 énoncés extraits de 52 dialogues. Les énoncés retenus pour le test sont ceux qui contiennent au moins un segment qui correspond à un arbre local ou global dans notre grammaire.

Pour tester le système nous avons choisi une méthode relativement simple qui est basée sur la distinction entre trois types d'erreurs : insérer, suppression ou substituer **un arbre élémentaire.** Le test a été fait sur les transcriptions.

Les résultats de notre évaluation sont présentés dans le tableau suivant :

Insertion	Substitution	Suppression	Total
1,7 %	1,8 %	11,4 %	14,9 %

Tableau 12. Résultat du test du système Safir

Le taux bas des segments insérés ou substitués est dû principalement au nombre relativement réduit d'ambiguïtés dans notre corpus ainsi qu'à la bonne désambiguïsation de la grammaire notamment grâce aux arbres globaux. Quant aux arbres supprimés, nous pouvons classer les raisons principales de ces erreurs dans deux groupes différents :

1. **Raisons en rapport avec les données :** une bonne partie des problèmes de suppression des arbres est due à la non-représentation de ces arbres dans notre corpus d'entraînement. Ces problèmes sont considérés comme secondaires dans la mesure où il faut avoir un corpus plus large pour les éviter.
2. **Raisons en rapport avec l'approche adoptée et l'état du système :** la majorité des cas d'échec de dépassement du bruit était essentiellement due à l'incomplétude de la stratégie de nettoyage. En général, les grammaires de nettoyage se sont montrées assez efficace au niveau des arbres globaux alors qu'elles ont manifesté certaines limitation au niveau des arbres lexicaux et locaux. En effet, l'une des principales limitations des grammaires de nettoyage est l'incapacité de ces grammaires à nettoyer un mot non pertinent localisé au sein d'un arbre local. Dans certains cas, ces règles ont même causé l'échec de l'analyse en faisant des fausses délimitations.

Ce test a été complété par une série de petits tests informels d'énoncés qui contiennent des cas difficiles non observés dans notre corpus de test formel pour avoir une idée sur le comportement du système dans ce genre de situations. Ces tests nous ont permis de constater, par exemple, la non-suffisance de la stratégie sélective pour le traitement de certaines formes d'extragrammaticalités comme les autocorrections qui nécessitent une considération plus fine. De même, le système a manifesté des incomplétudes importantes dans le traitement de certains phénomènes syntaxiques complexes comme la négation. Par contre, les résultats avec des énoncés qui contiennent des ellipses se sont révélées assez positives.

2.1.6 Bilan général du système Safir

Voici un bilan général des aspects clés du système Safir :

1. **Le prétraitement :** le prétraitement a une fonction très limitée au sein du système SAFIR qui ne dépasse pas l'adaptation du format de l'entrée aux contraintes de l'analyseur. Une extension possible de ce module consiste en l'ajout d'un filtre qui supprime tous les mots qui ne font pas partie du lexique du système. Cela permet de résoudre l'un des principaux problèmes du système qui est l'insertion de mots inconnus au sein d'un îlot pertinent causant ainsi l'échec du système à analyser cet îlot. Ce problème nécessite plus d'investigations dans le futur notamment en ce qui concerne l'emplacement de ce filtre au sein de l'architecture (si nous avons besoin de l'information : existence de mots inconnus pour un traitement quelconque) ou s'il existe des mots externes au lexique que le système doit traiter comme des noms propres, etc.
2. **Le formalisme :** le formalisme utilisé s'est révélé bien adapté pour le traitement des énoncés avec des phénomènes fréquents à l'oral comme les ellipses et il présente aussi l'avantage d'intégrer des informations fournies par la tâche, source d'informations assez fiable dans le contexte d'un système de dialogue. Cela renforce généralement la robustesse de l'analyse. Cependant ce formalisme semble assez limité pour le traitement de certains phénomènes qui

nécessitent l'intégration de la syntaxe de manière déclarative comme la négation, les modifieurs en général et la coordination.

3. **L'algorithme d'analyse :** les RTRs ainsi que la stratégie sélective et l'approche d'analyse partielle semble bien adaptées à la tâche de l'analyse d'énoncés spontanés. Cependant, malgré son adaptation globale pour le traitement, la stratégie sélective s'est montrée parfois incapable d'ignorer les zones non pertinentes dans certains cas et elle a posé des problèmes de surgénération dans d'autres. Cela nécessite non le rejet de la stratégie sélective comme idée (puisque dans cet état-là nous estimons que son apport est supérieur aux erreurs qu'elle cause) mais plutôt son amélioration afin de maximiser ses avantages et réduire ses inconvénients.
4. **Besoin d'un dispositif spécifique pour le traitement des extragrammaticalités :** nos tests informels ont confirmé nos idées selon lesquelles l'approche sélective toute seule n'est pas suffisante pour le traitement propre des extragrammaticalités et en particulier celles qui ont un effet sur l'interprétation sémantique comme l'autocorrection. Cela nécessite l'intégration d'un module spécifique dans notre système qui joue un rôle similaire à celui de Corrector[44].

2.2 La solution des problèmes de Safir : le système Oasis

Après notre expérience encourageante avec le système SAFIR, nous avons décidé de construire un système qui intègre les points positifs de Safir avec l'amélioration de ses points de faiblesse. Ainsi, nous avons développé le système Oasis. Ce nouveau système est basé sur le formalisme Sm-TAG et il intègre, entre autres, un module de traitement des extragrammaticalités basé sur notre travail dans le système Corrector.

2.2.1 Les requis du système Oasis

Les requis du système Oasis sont similaires à ceux du système Safir mais s'en distinguent par les points suivants :

1. L'entrée du système Oasis est la sortie d'un système de reconnaissance de la parole.
2. Etant réalisé dans le cadre d'un système de traduction automatique de la parole, Oasis doit fournir une analyse fine.
3. Le domaine de l'application est plus large que celui de Safir.

2.2.2 Architecture du système Oasis

Le traitement dans Oasis se fait selon trois étapes principales :

- L'étiquetage.
- L'analyse syntactico-sémantique et le post-traitement.

---

[44] Nous aimerons attirer l'attention du lecteur que la réalisation du système SAFIR est antérieure à celle de CORRECTOR chronologiquement et donc ce constat était aussi l'un des principaux motifs derrière notre investigation des extragrammaticalités et leur normalisation.

Le schéma général de l'architecture d'Oasis est présenté dans la figure suivante :

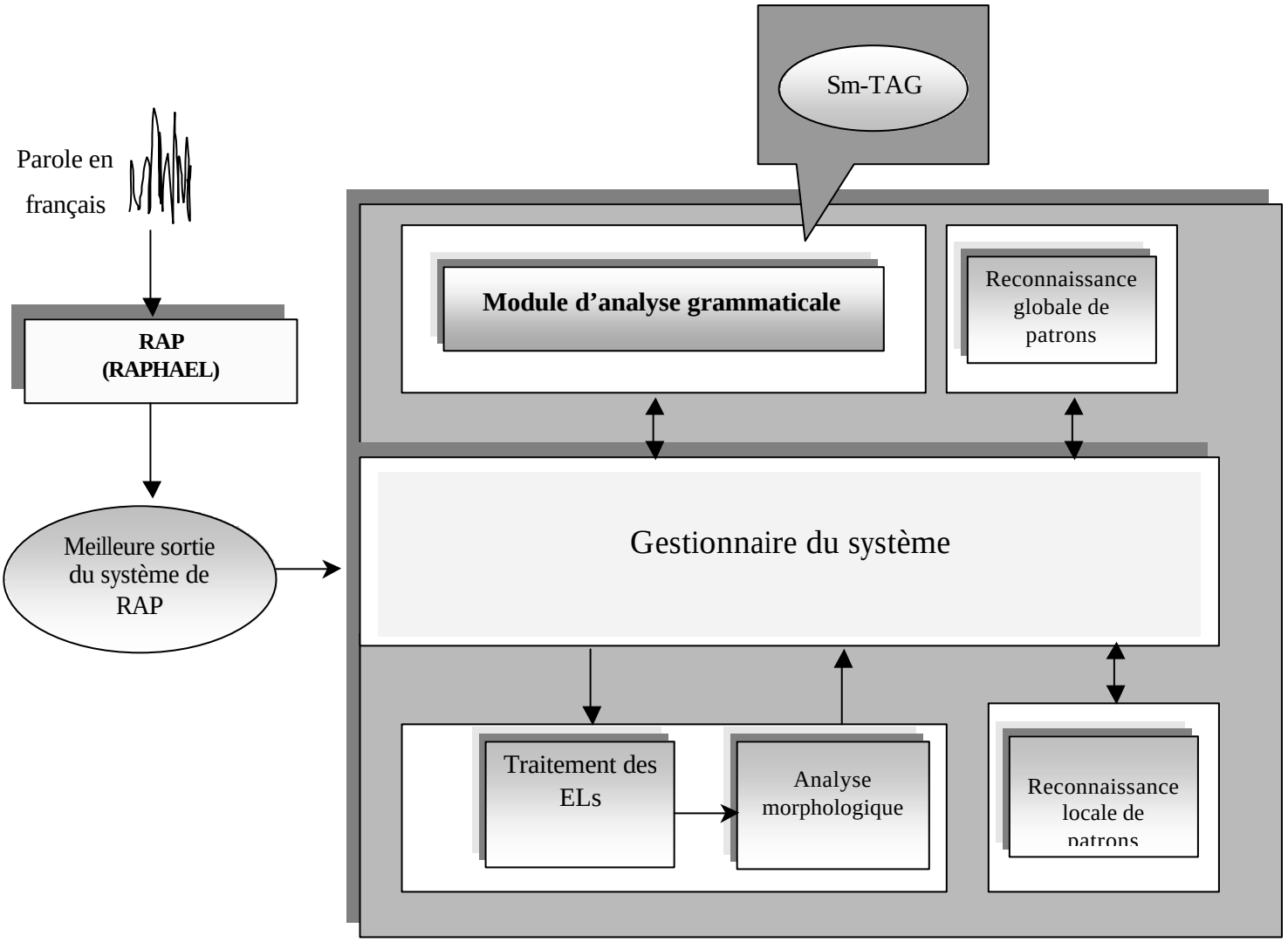


Figure 77. Architecture générale du système Oasis

Comme nous pouvons le constater dans la figure précédente, le système Oasis est basé sur la répartition du traitement à un ensemble de modules hétérogènes qui communiquent à travers un module central similaire à celui que nous avons utilisé avec le système Corrector.

2.1.1.75 Le gestionnaire de système

Ce module est similaire au gestionnaire de système que nous avons utilisé dans le système Corrector. En effet, il s'agit d'une unité dont la fonction est la transmission de l'information entre les différents modules. Le flux de l'information résultant de l'interaction des différents modules via le gestionnaire du système est présenté dans le tableau suivant :

Entrée	Source	Destination	Traitement	Destination
Meilleure hypothèse de reconnaissance	Raphael	GS		
Meilleure hypothèse de reconnaissance	GS	Ttraitement lexical	Traitement des extragrammaticalités lexicales, normalisation de certains mots oraux et analyse morphologique	GS
Enoncé oral analysé morphologiquement et dont les phénomènes lexicaux sont normalisés	GS	Reconnaissance locale de patrons	Traitement des répétitions et des autocorrections dont l'étendue est limitée	GS
Enoncé dont les répétitions et les auto-corrections locales sont traitées	GS	Reconnaissance globale de patrons	Traitement des répétitions et des autocorrections dont l'étendue est large	GS
Enoncé dont les répétitions et les autocorrections sont traitées	GS		Analyse grammaticale avec le formalisme Sm-TAG	GS
Analyse grammaticale de la meilleure hypothèse de reconnaissance	GS	Interface utilisateur		

Tableau 13. Le flux de l'information dans le système Oasis

Tout comme dans Corrector, le gestionnaire de système est aussi chargé de l'adaptation du format de l'information pour l'entrée de chaque module.

2.1.1.76 Le module de reconnaissance

L'entrée d'Oasis est la sortie du système de reconnaissance RAPHAEL préparé au sein de l'équipe GEOD du laboratoire CLIPS-IMAG. RAPHAEL a été construit sur la plate-forme du système Janus de l'ISL-CMU tout d'abord par Mohamed AKBAR (Akbar et Caelen, 1998) et ensuite par Dominique Vaufreydaz (Vaufreydaz *et al.*, 1999), (Vaufreydaz *et al.*, 2000). Il s'agit d'un système indépendant du locuteur de vocabulaire moyen. Il est composé de deux modules principaux :

1. Un modèle acoustique : il s'agit d'un modèle markovien entraîné sur le corpus BREF-80. Ce corpus contient 12 heures de parole continue de 72 locuteurs et un vocabulaire d'environ 5500 variantes phonétiques de 2900 mots (Lamel *et al.*, 1991).
2. Le modèle de langage: RAPHAEL utilise un modèle à base de classe qui permet d'ajouter des noms propres facilement au vocabulaire. Ce modèle a été entraîné sur un corpus d'environ 10 gigabytes de documents texte et HTML. Ces données ont été collectées de l'espace français de l'Internet et ont été adaptées à l'aide d'une série de prétraitements spécifiques aux contraintes de modèles de langages (comme la suppression des tags de HTML) (Vaufreydaz *et al.*, 2000). Ce modèle général a ensuite été optimisé pour la tâche de réservation touristique.

Comme première intégration, nous avons jugé bon de commencer par le traitement de la meilleure hypothèse de reconnaissance de la parole. Ce choix est motivé par la simplicité de ce mode d'intégration ainsi que la bonne qualité de reconnaissance possible avec le système RAPHAEL (cela réduit l'intérêt d'utilisation de connaissances linguistiques pour l'amélioration des résultats de reconnaissance).

2.1.1.77 Le prétraitement

Les modules de prétraitement sont destinés à préparer l'entrée de manière à rendre son traitement plus facile par les analyseurs syntaxiques et sémantiques.

Deux phases principales se distinguent au sein du prétraitement : le traitement lexical et le traitement supralexicale.

2.2.2.1.1 Le traitement lexical

Le traitement lexical se fait selon deux étapes principales :

1. **Filtrage des mots inconnus[45] :** l'une des principales limitations de la stratégie sélective que nous avons observée dans le système Safir est que cette stratégie est uniquement opérationnelle entre les segments en cours d'analyse mais pas au sein de chaque segment. Pour limiter l'effet de ce problème, nous avons décidé d'ajouter un module de filtrage qui supprime les mots externes au lexique *a priori*. Cela permet de réduire les cas d'échec d'analyse dus à des mots

[45] Dans le contexte d'un système d'analyse linguistique intégré avec un module de reconnaissance de la parole, les mots inconnus sont réduits à ceux qui font partie du lexique du module de reconnaissance mais pas de celui du système d'analyse linguistique.

inconnus qui s'insèrent au sein d'un segment pertinent et qui cause par conséquent l'échec de l'analyse de ce segment.

2. **Traitement des extragrammaticalités lexicales et des phénomènes lexicaux oraux :** le traitement des extragrammaticalités lexicales consiste à convertir les formes orales des mots en leur versions écrites *standards*. Ces extragrammaticalités sont (d'après notre observation informelle) moins fréquentes en français qu'en anglais américain. En effet, les amalgames couramment utilisés en anglais oral aux Etats-Unis ne sont pas aussi systématiques en français oral. Ce que nous observons principalement ce sont des simplifications phonétiques comme: *ch'ui* pour *je suis* ou *ouais* pour *oui.* Ces phénomènes ont été pris en considération dans la version du système qui est destinée à traiter les transcriptions en enrichissant le lexique des formes de l'oral. Dans la version actuelle, qui a pour entrée la sortie de RAPHAEL, nous avons uniquement les formes standards en entrée du système et nous n'avons pas de problèmes particuliers à cause de ces phénomènes.

2.2.2.1.2 Analyse morphologique

Nous utilisons le dictionnaire de notre système avec un nombre restreint de règles morphologiques pour désambiguïser les items lexicaux. En effet, les ambiguïtés lexicales observées dans notre corpus sont très limitées puisque le nombre du lexique de notre tâche n'est pas très élevé. Ainsi, les confusions entre les mots ne sont pas très fréquentes. Par exemple, le mot *réserve* peut être associé à deux catégories morphologiques en même temps : verbe et nom. Le nom *réserve* n'étant pas observé dans notre corpus, il n'est pas ajouté au lexique et l'ambiguïté morphologique n'est pas perçue par le système.

Par ailleurs, les conditions de traitement peuvent créer des ambiguïtés artificielles qui ne sont pas observées dans les contextes de systèmes de traitement de l'écrit. Par exemple, le système de reconnaissance peut sortir *a* ou *à* correspondant à la préposition ou au verbe. Pour éviter ce genre de cas, nous avons équipé la grammaire d'entrées lexicales correspondants à ces deux formes (c'est-à-dire, chacune de ces deux formes est associée aux deux catégories morphologiques).

2.1.1.78 Traitement des extragrammaticalités supralexicales

La partie principale du module de prétraitement que nous utilisons ici est une adaptation au français du module d'analyse par patrons que nous avons développé pour l'anglais. Les motivations de cette adaptation sont les suivantes :

1. Les structures de répétitions et d'autocorrections sont pratiquement les mêmes dans toutes les langues.
2. Cela nous évite de refaire le même travail que nous avons effectué sur l'anglais surtout que les corpus de dialogues oraux spontanés correspondants à notre tâche ne sont pas très disponibles en français.

L'adaptation n'a pas été faite sur des critères purement personnels. En effet, cela a été fait en analysant un mini corpus de 80 cas d'extragrammaticalités supralexicales extraits du corpus de dialogues spontanés collectés récemment dans l'équipe GEOD-CLIPS dans le cadre du projet Nespole. Les principales modifications apportées après son adaptation sont résumées dans les points suivants :

- Adaptation des parties des patrons qui correspondent aux zones d'édition en implantant des règles sémantiques pour analyser des expressions comme: *attendez une minute s'il vous plaît, ne quittez pas, enfin*, etc. Les expressions ajoutées étant à la fois celles observées dans le mini corpus français ou des traductions que nous avons effectuées des expressions anglaises observées.
- Adaptation des patrons à la sortie de l'analyseur morphologique dont les étiquettes sont différentes de celle du tagger de Xerox ainsi que la modification des équivalences des catégories pour les autocorrections.

2.1.1.79 La grammaire

Deux sources d'informations ont été utilisées pour écrire la grammaire Sm-TAG du système Oasis :

1. **La grammaire S-TSG :** vu les similarités entre le formalisme Sm-TAG et le formalisme S-TSG que nous avons utilisé dans SAFIR, la première étape de notre travail a consisté à convertir les arbres de la S-TSG en arbres Sm-TAG. Cette conversion a été faite selon deux procédures :
   i. L'adoption des arbres qui remplissent les conditions de bonne formation de la Sm-TAG. En effet, nous avons trouvé que certains arbres de la S-TSG (en particulier les arbres globaux qui sont les moins contraints dans la Sm-TAG) correspondent tel qu'ils sont à des arbres de la Sm-TAG.
   ii. Modification des arbres S-TSG qui ne remplissent pas les conditions des arbres de la Sm-TAG. Cela a été fait essentiellement pour convertir les arbres lexicaux en y ajoutant un nœud supplémentaire ou en changeant l'un de ses non-terminaux. Par ailleurs, nous avons effectué certaines modifications sur les arbres locaux en divisant certains arbres de la S-TSG en deux ou au contraire en unissant deux arbres locaux différents.
2. **L'analyse directe de corpus :** cette analyse de corpus est faite pour compléter l'information que nous avons obtenue des règles de la S-TSG d'une part et d'autre part pour élargir la couverture de notre grammaire (la grammaire S-TSG a été conçue pour un prototype élémentaire et ne couvre que les concepts simples). Le déroulement de l'analyse du corpus est similaire à celui que nous avons décrit pour la S-TSG.

Ainsi, la grammaire Sm-TAG écrite contient au total 1480 arbres dont 211 arbres locaux et globaux et 1269 arbres lexicaux.

2.1.1.80 L'algorithme d'analyse

L'algorithme que nous avons adopté pour l'analyse avec la Sm-TAG est un algorithme à deux passes : une passe pour l'analyse syntaxique et une passe pour l'analyse sémantique. La raison principale de ce choix est d'augmenter la rapidité du traitement. En effet, l'utilisation de plusieurs passes qui s'appliquent en cascade est une approche qui a été adopté dans différents travaux (Abney, 1995), (Aït-Mokhtar et Chanod, 1997) pour réduire la combinatoire due à l'interaction des différents nivaux d'analyse. Par ailleurs, sur le plan linguistique, les opérations d'association peuvent être vues comme un moyen pour construire un noyau syntaxique local sur la base duquel se construit une représentation sémantique globale avec l'opération de substitution.

2.2.2.1.3 La première passe

La première passe consiste à construire les noyaux syntaxiques locaux sur la base desquels le niveau sémantique sera construit.

1. **Description générale de l'algorithme :** l'objectif de l'algorithme de la première passe est de construire un premier noyau syntaxique sur la base duquel se construit le niveau sémantique. Ce noyau consiste en un ensemble d'arbres intermédiaires (des arbres d'analyse dont la racine n'est pas le non-terminal distingué) qui ne sont pas connectés aussi bien que des mots non-analysés. Ainsi, la fonction de cette première passe est de détecter les arbres lexicaux auxiliaires et de les associer aux arbres lexicaux initiaux appropriés et puis d'effectuer l'opération de propagation sur l'arbre intermédiaire obtenu. L'approche générale de l'algorithme d'analyse que nous avons adopté est inspirée par l'algorithme du type Early.
2. **Notation :** voici les éléments de base que nous avons adoptés dans notre présentaiton de l'algorithme :
   - La grammaire Sm-TAG : $G = (\sum, NT, I, A)$ (voir la deuxième partie de cette thèse pour la définition formelle d'une grammaire Sm-TAG).
   - Les lettres grecques m,n et r sont utilisées pour désigner les nœuds des arbres élémentaires. Chacun de ces nœuds est associé à la catégorie syntaxique qui le décore. Par exemple les deux éléments suivants : m_A,n_B montrent que le nœud m est décoré par le non-terminal A et que le nœud$n$ est décoré par le non-terminal B.
   - Les arbres sont représentés avec un format inspiré des règles de réécriture des CFGs. Ainsi, le non-terminal le plus à gauche dans la règle correspond à la racine de l'arbre et les terminaux correspondent aux nœuds feuilles de l'arbre. Les parenthèses sont utilisés pour représenter les niveaux hiérarchiques dans les arbres. Par exemple, dans la règle suivante $m_A \rightarrow \cdot(m_B \rightarrow m_c)\ (m_D \rightarrow m_e)$ les nœuds m_B et m_d sont les nœuds fils la racine de l'arbre : m_A. Ainsi, l'arbre correspondant à cette règle a la forme suivante :

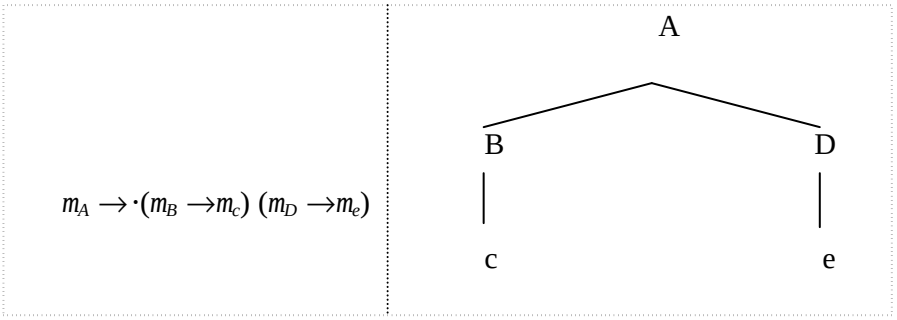


Figure 78. Un arbre d'analyse et son équivalence selon le format que nous avons proposé

Finalement, nous utilisons une lettre grecque dans la partie droite d'une règle (comme : $m_A \rightarrow \alpha$) pour désigner les séquences de k nœuds fils où $k \geq 0$.

- Le prédicat *LeftAux*(r_A) est vrai si et seulement r_A est la racine d'un arbre auxilliaire gauche.
- Le prédicat *RightAux*(r_A) est vrai si et seulement r_A est la racine d'un arbre auxiliaire droit.
- Le prédicat *Init*(r_A) est vrai si et seulement si r_A est la racine d'un arbre initial.
- Soit la chaîne d'entrée $W : w_1 \ldots w_n$, l'unité de base de l'algorithme a la forme suivante : $[m_A \rightarrow \cdot\alpha,\ i]$ où $1 \leq i \leq n$. Cette unité veut dire que la racine de l'arbre couvre un item lexical situé dans le point i du chart : notons que nous avons utilisé un seul indice spatial plutôt que deux (comme c'est le cas dans la plupart des autres approches d'analyse tabulaire (voir par exemple (Shabes et Waters, 1944)) étant donné que la couverture des arbres lexicaux dans la Sm-TAG est limitée à un seul item. Finalement, il n'est probablement pas inutile de rappeler que le point $\cdot$ permet de séparer les nœuds fils qui ont été parcourus par l'algorithme (ces nœuds sont situés à gauche du point) des nœuds fils qui ne l'ont pas été encore (ils sont situés à droite du point).

3. **L'algorithme :**

Initialisation :

$$\frac{Init(m_s)}{[m_s \rightarrow \cdot\alpha,\ 0]}$$

Objectif :

$$\frac{Init(m_s)}{[m_s \rightarrow \alpha\cdot,\ n]}$$

Association gauche :

$$\frac{[m_A \rightarrow \cdot\alpha,\ i]}{[r_A \rightarrow \cdot\gamma,\ i]} \quad \text{LeftAux}(r_A)$$

$$\frac{[m_A \rightarrow \cdot\alpha,\ i] \quad [r_A \rightarrow \cdot\gamma,\ i+1]}{[m_A \rightarrow \cdot\alpha,\ i+1]} \quad \text{LeftAux}(r_A)$$

Scan :

$$\frac{[m_A \rightarrow \alpha\cdot n_a b,\ i]}{[r_A \rightarrow \cdot\alpha n_a \cdot b,,\ i+1]} \quad a = a_{i+1}$$

Association droite :

$$\frac{[m_A \rightarrow \cdot\alpha,\ i]}{[r_A \rightarrow \cdot\gamma,\ i]} \quad \text{RightAux}(r_A)$$

$$\frac{[m_A \rightarrow \cdot\alpha,\ i] \quad [r_A \rightarrow \cdot\gamma,\ i+1]}{[m_A \rightarrow \cdot\alpha,\ i+1]} \quad \text{RightAux}(r_A)$$

Propagation inductive :

$$\frac{[m_A \rightarrow \cdot(m_B \rightarrow m_c)\ (m_D \rightarrow m_e),\ i]}{[m_D \rightarrow \cdot(m_A \rightarrow m_c)\ (m_D \rightarrow m_e),\ i]} \quad \text{LeftAux}(m_B)$$

$$\frac{[m_A \rightarrow \cdot(m_B \rightarrow m_c)\ (m_D \rightarrow m_e),\ i]}{[m_B \rightarrow \cdot(m_A \rightarrow m_c)\ (m_D \rightarrow m_e),\ i]} \quad \text{RightAux}(m_D)$$

Figure 79. La première passe de l'algorithme d'analyse de la Sm-TAG

i. Le premier item de l'algorithme permet d'initialiser le chart en y ajoutant toutes les règles du type $[m_s \rightarrow \cdot\alpha, 0]$ où m_s est la racine d'un arbre élémentaire quelconque.

ii. La clause d'arrêt veut dire que l'analyse est satisfaite si tous les éléments de l'entrée sont parcourus et si la racine de l'arbre obtenu correspond à celle d'un arbre élémentaire.

iii. **Le scan :** la règle de scan permet de détecter et de consommer les terminaux dans la chaîne d'entrée.

iv. **L'association gauche et droite :** cette étape consiste à associer les arbres lexicaux auxiliaires aux arbres lexicaux initiaux correspondants. Voici, à titre d'exemple, le schéma de l'opération d'association simple gauche :

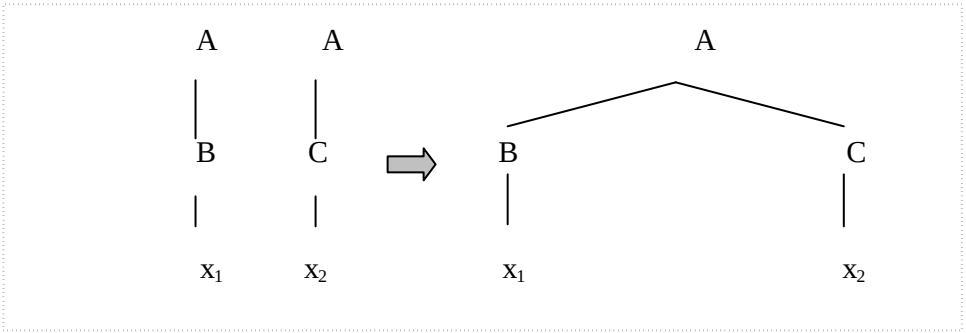


Figure 80. Schéma de l'opération d'association simple gauche ainsi que la règle d'inférence utilisée pour son implantation

Comme nous pouvons le remarquer, la sortie de cette opération est un arbre intermédiaire dont la racine et les ancres sont respectivement la racine commune des deux arbres lexicaux qui le forment et leurs ancres. Cet arbre se combine avec le reste des arbres de la grammaire avec l'opération de substitution.

v. La propagation gauche et la propagation droite : l'objectif principal de cette étape est d'adapter les arbres intermédiaires obtenus dans l'étape précédente. A titre d'exemple, voici le schéma général de la propagation inductive gauche :

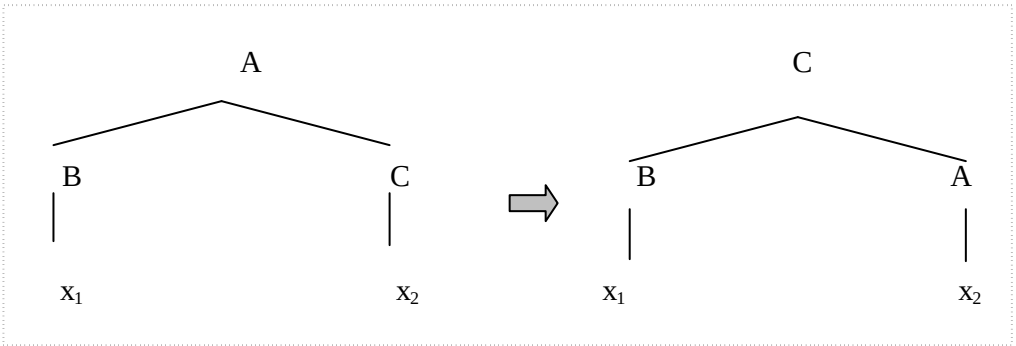


Figure 81. Le schéma général de l'opération de propagation inductive gauche

Comme nous pouvons le remarquer, les règles d'inférence permettent de propager la racine de l'arbre initial et en faire la racine de l'arbre intermédiaire.

4. **Complexité de l'algorithme :** comme nous pouvons le remarquer dans notre algorithme, la règle d'inférence la plus complexe que nous avons utilisée contient deux variables spatiales seulement. Ainsi, nous pouvons dire que la complexité temporelle de l'algorithme est linéaire au pire des cas : $O(n^2)$ où n est la longueur de l'entrée. Notons que cette réduction du nombre de variables est dû à la limitation de couverture des arbres lexicaux à un seul item lexical ce qui nous a permis de pouvoir utiliser une seule variable spatiale pour indiquer la couverture de ces arbres.

2.2.2.1.4 La deuxième passe

Nous utilisons l'opération de substitution afin de lier les arbres intermédiaires obtenus dans la première passe d'analyse ainsi que les mots non-analysés dans cette passe. Pour effectuer l'opération de substitution, nous utilisons les RTRs. Outre que les avantages des RTRs que nous avons présentés dans le chapitre précédent, le choix des RTRs a plusieurs motivations dont les principales sont résumées dans les deux points suivants :

- Bien que la Sm-TAG est faiblement équivalente à une CFG, nous avons vu que tous les arbres élémentaires qui se combinent avec l'opération de substitution peuvent être remplacés par des règles de réécriture équivalente (voir le deuxième chapitre de la troisième partie de cette thèse ainsi que (Schabes et Waters, 1995) pour plus de détails sur ce point). Ainsi, nous pouvons représenter tous les arbres du formalisme Sm-TAG qui se combinent avec l'opération de substitution comme des RTRs sans perdre de l'information.
- La conversion des grammaires d'arbres en automates est une approche qui a été adoptée par d'autres chercheurs comme (Lopez, 1999a). Les motivations d'un tel choix sont : la meilleure connaissance des propriétés computationnelles des automates que celles des

arbres ainsi que la bonne visualisation des données avec les automates notamment grâce à leur aspect séquentiel.

Les RTRs que nous avons utilisés sont enrichis avec deux propriétés qui les rendent plus adaptés au traitement du langage oral : l'analyse partielle et la stratégie sélective.

1. **L'analyse partielle :** l'approche d'analyse partielle que nous avons adoptée dans le système Oasis est similaire à celle que nous avons utilisée dans le système Safir : le système essaie d'abord de trouver une analyse qui maximise la couverture et lorsque cela est impossible il passe à des analyses partielles qui couvrent des segments de l'entrée plutôt que sa totalité.
2. **La stratégie sélective :** la stratégie sélective consiste à ignorer tous les mots considérés non pertinents pour la tâche. Nous avons vu que la première étape de cette sélection commence au prétraitement avec le filtrage des mots inconnus. Malgré son utilité, ce filtrage n'est pas suffisant pour le traitement. En effet, une bonne partie des problèmes d'échec d'analyse peut être due à des mots qui font partie du lexique mais qui ne sont pas dans un endroit qui permet au système de les traiter. Ainsi, nous avons décidé d'enrichir l'algorithme d'analyse avec une stratégie sélective. La solution que nous avons adoptée finalement consiste en la combinaison de deux techniques :

vi. **Les grammaires de nettoyage :** au niveau des arbres globaux, nous avons utilisé des grammaires de nettoyages similaires à celles utilisées dans le système Safir. L'utilisation de ces grammaires donne au système plus de souplesse en permettant à des segments non pertinents de séparer deux arbres (locaux ou globaux).

vii. **La fonction sélective :** au niveau des arbres locaux nous avons décidé d'utiliser une nouvelle version de la stratégie sélective étant donné que les grammaires de nettoyage ne se sont pas montrées complètement satisfaisantes sur ce niveau. Ainsi, nous avons proposé une solution simple basée sur la combinaison d'un arbre négatif à l'aspect descendant de notre approche. L'arbre négatif consiste en un arbre qui accepte toutes les unités qui ne sont pas acceptées comme un arbre bien formé dans la grammaire. La priorité de l'arbre négatif est la moins importante dans l'analyse. Ainsi, le système avant d'ignorer un mot de l'entrée il vérifie toutes les possibilités d'analyse de ce mot. Cela attribue à notre approche tous les avantages d'une approche d'analyse complète avec la souplesse des approches sélectives. Outre la localisation des zones pertinentes dans l'entrée, la stratégie sélective a pour fonction de traiter certaines formes d'extragrammaticalités jouant ainsi le rôle de la deuxième muraille de défense contre les extragrammaticalités. En effet, les cas qui peuvent être traités par la stratégie sélective sont :

- Extragrammaticalités qui apparaissent dans les zones non pertinentes : répétitions ou autocorrection de mots ou de séries de mots non pertinents.

- Toutes les extragrammaticalités qui impliquent un segment pertinent inférieur à un arbre local. Ainsi, un mot pertinent qui ne forme pas tout seul un arbre local est considéré comme non pertinent s'il est répété deux fois et par conséquent ce mot est ignoré par la stratégie sélective. Pour mettre au clair ce point prenons l'exemple suivant :

  Je je voudrais une chambre (106)

  Comme nous pouvons le constater dans l'énoncé précédent, le premier *je* est parfaitement pertinent par rapport à la tâche et il fait partie du lexique du système mais il est ignoré étant donné qu'il ne constitue pas un arbre local entier. La deuxième occurrence est traitée comme une occurrence normale d'un mot pertinent.

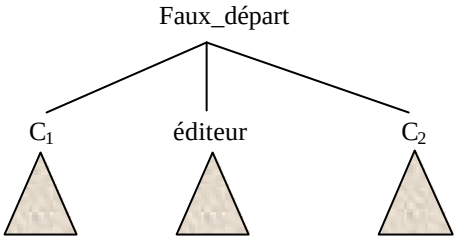
2.1.1.81 Le post-traitement

La fonction principale du module de post-traitement est de normaliser les faux-départs et les incomplétudes dans les énoncés de l'entrée. La spécificité principale de ce module par rapport au module implanté pour l'anglais est que ce module est basé sur les informations sémantiques fournies par l'analyseur basé sur la Sm-TAG et pas sur les informations purement syntaxiques comme dans le cas de l'analyseur de l'anglais. En effet, le principe de base de ce module est de détecter les faux-départs sur la base des anomalies sémantiques. Cela est fait par un ensemble de méta-règles sémantiques dont les principales contraintes sont les suivantes :

1- L'ordre des arbres sémantiques dans un énoncé.

2- La racine de l'arbre ainsi que la catégorie fonctionnelle associée à cette racine qui peut être: acte de parole, concept, argument.

3- Informations sur la zone d'édition qui peut séparer les arbres impliqués dans l'extragrammaticalité.



Transition_ impossible (C_1, C_2).

Figure 82. Le schéma général des méta-règles utilisées dans le traitement des incomplétudes et des faux-départs

Si la transition entre les arbres C_1 et C_2 est impossible, alors le système décide que l'énoncé en cours d'analyse contient un faux-départ. Dans ce cas, le système supprime l'arbre C_1 avec la zone d'édition

si celle-ci est considérée comme étant non nécessaire au traitement de C_2. Pour mettre au clair ces règles, examinons l'exemple suivant :

C'est pour euh Je voudrais une réservation pour deux personnes. (107)

Dans l'énoncé précédent, nous remarquons que nous avons un faux départ qui consiste en deux segments qui correspondent à une formule de demande: *c'est pour* et *je voudrais*. Ces deux segments ne sont pas normalisés par le module de prétraitement étant complètement différents sur le plan de leur forme. De même, ces deux segments étant parfaitement bien formés, ils sont analysés par le système comme deux segments indépendants auxquels le système associe la catégorie sémantique *formule_de_demande*. Pour résoudre ce faux départ, le système examine les deux catégories sémantiques associées aux deux segments et décide qu'il s'agit d'un faux départ étant donné que la succession de deux catégories formules de demande est impossible.

A ce stade du développement, comme nous ne disposons pas d'un nombre suffisant d'incomplétudes et de faux départs en français, notre objectif principal est de montrer que nous pouvons faire des traitements spécifiques pour les faux-départs et les incomplétudes. Sur le plan pratique, cela nous permet d'avoir une première évaluation de cette approche à travers les différentes évaluations que nous avons l'intention de faire. Ainsi, un noyau de cinq méta-règles est implanté dans cette version du système.

2.1.1.82 Discussion de l'architecture d'Oasis

La conception de l'architecture d'Oasis a été faite selon un nombre de considérations dont les principales sont :

1. **Considérations générales :** comme nous avons vu avec le système Corrector, le gestionnaire de système à base de Hub a pour fonction de transmettre l'information :
   i. Indépendance des sources de connaissance : les trois principaux blocks pour le traitement des extragrammaticalités (l'analyse morphologique, le traitement des extragrammaticalités et l'analyse grammaticale) sont assez indépendants les uns des autres et ne nécessitent pas une interaction avancée entre les modules. Ceci est dû principalement à l'intégration de différentes sources de connaissances qui nécessitent des interactions complexes dans le cadre de la Sm-TAG.
   ii. La neutralité applicative et la limitation du système au niveau de l'analyse linguistique réduisent elles aussi les possibilités et les besoins d'interactions entre les modules. Par exemple, nous ne disposons pas d'un module de niveau supérieur comme un gestionnaire de dialogue qui émet des attentes qui guident le module d'analyse.
2. **Considérations logicielles :** d'un point de vue logiciel, notre architecture consiste en un ensemble de moldules qui joue le rôle de serveur à une unité centrale qui, à son tour, joue le rôle

de client. Le raisons pour lesquelles nous avons utilisé un gestionnaire de système sont assez similaires à celles que nous avons donné pour le système Corrector :

i. **Hétérogénéité des soruces de connaissances à intégrer :** comme nous avons vu, notre architecture intègre des modules dont les domaines sont assez variés : analyse morphologique, traitement des extragrammaticalités et analyse grammatical. L'utilisation d'un module indépendant de la tâche comme un espace commun où les différents modules peuvent communiquer facilite l'interaction de ces modules puisque nous n'avons pas à considérer la nature des modules pour les faire communiquer.

ii. **Souplesse :** comme les modules ne communiquent pas directement, il est relativement facile d'ajouter un nouveau module ou de remplacer un module existant par un autre (En cas de besoin de comparaison entre différentes techniques par exemple). Pour ce faire, il suffit de remplacer l'appel à l'ancien module par celui du nouveau module et, en cas de besoin, de mettre à jour le dispositif de formatage des données (qui adapte le format de l'entrée aux contraintes du module suivant et qui adapte le format de sa sortie aux contraintes du module d'après) à l'entrée et à la sortie de ce module.

iii. **Portabilité :** la modularité de l'approche rend possible la réutilisation de certains modules dans différentes applications y compris le gestionnaire de système.

2.2.3 Implantation du système Oasis

Tout comme les systèmes Corrector et Safir, le système Oasis a été implanté en utilisant PROLOG. Le système est composé de 6 fichiers dont les noms et les tailles sont présentés dans le tableau suivant :

Fichier	Nombre des lignes
Main_Oasis	448
Main_parsing_module	4566
Pattern_preprocessing	2211
Association_module	160
Tree_drawer	534
Induction_rules	460
Total	8379

Tableau 14. Présentation de l'organisation générale du code du système Oasis en fichiers

Comme nous pouvons le constater dans le tableau précédent, le programme a été divisé en fichiers selon des fonctionnalités spécifiques : cela peut être un module particulier comme le module de

traitement par patrons : pattern_preprocessing ou une opération indépendante comme l'opération d'association : association_module.

2.2.4 Evaluation du système Oasis

L'objectif de notre évaluation est de montrer les avantages et les limites de notre approche afin de situer l'efficacité de notre système dans le contexte des travaux existants. Pour ce faire, nous avons décidé d'effectuer trois évaluations qui sont à la fois différentes et complémentaires. La première de ces évaluations porte sur le calcul de la complexité effective de l'algorithme d'analyse alors que les deux autres portent sur l'efficacité du système en terme de traitement. En effet, il s'agit d'une évaluation quantitative et d'une évaluation qualitative. L'objectif de l'évaluation quantitative est de montrer l'état d'achèvement de l'implantation en terme de couverture lexicale et grammaticale alors que l'évaluation qualitative tente d'aller plus loin en diagnostiquant les raisons d'échec et de réussite d'analyse et en les liant à l'approche utilisée.

2.1.1.83 Evaluation du temps de calcul de notre algorithme d'analyse

Afin d'évaluer le temps de calcul de notre algorithme, nous avons choisi un corpus de 588 énoncés. Les énoncés choisis sont extraits du corpus de réservation hôtelière de même que du corpus collecté dans la campagne d'évaluation par défi que nous allons présenter plus loin. Les fréquences des énoncés utilisés par rapport à leurs longueurs sont présentées dans le graphe suivant :

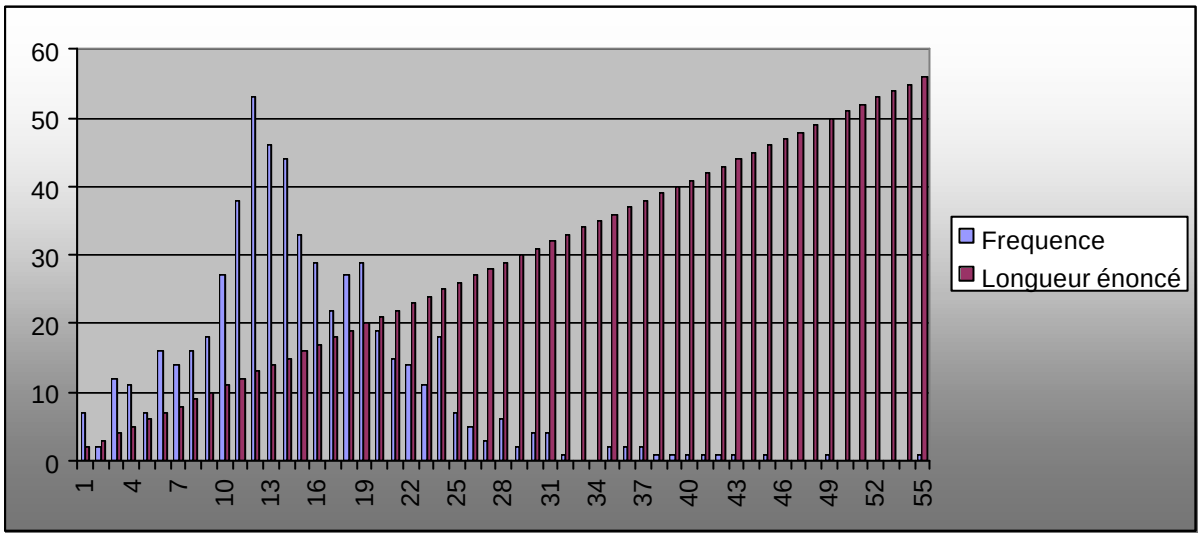


Figure 83. Les longueurs des énoncés utilisés pour le test comparées à leurs fréquences

Comme nous pouvons le remarquer dans le graphe précédent, les fréquences les plus importantes sont situées dans la zone entre 7 et 23 mots avec un sommet au milieu (53 occurrences des énoncés de 13 mots).

Le résultat de l'analyse des temps de calcul sur tout le corpus est présenté dans le graphe suivant :

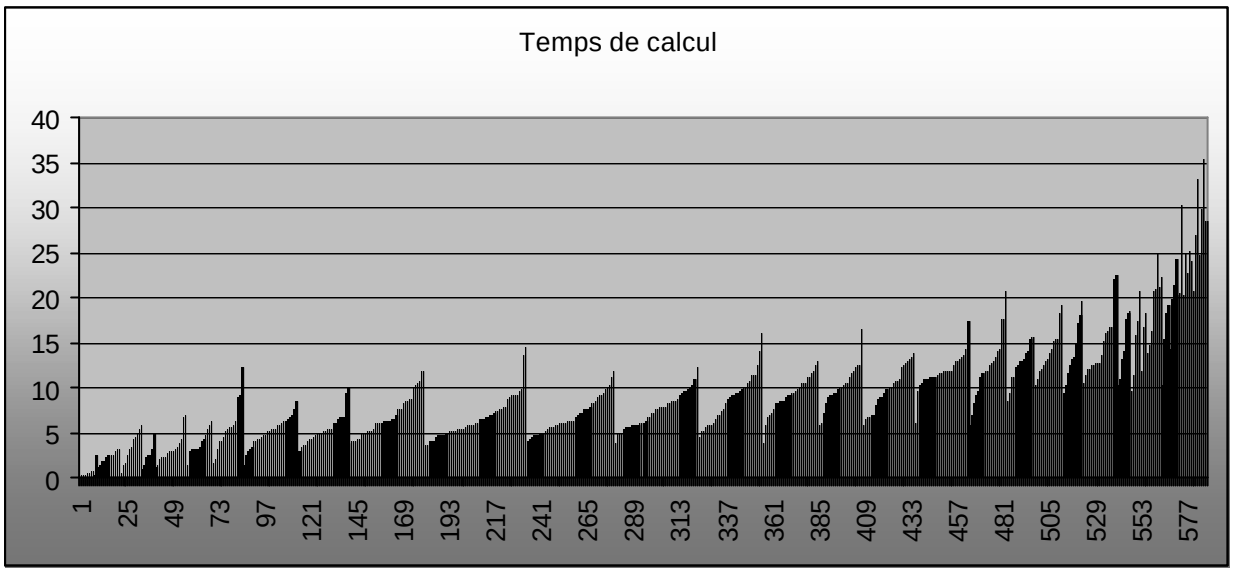


Figure 84. Les temps de calcul obtenus sur la totalité du corpus de test

Comme nous pouvons le remarquer dans le graphe précédent, l'augmentation tend à être linéaire entre les différents ensembles d'énoncés de même longueur. Pour avoir une idée plus claire de la courbe du temps de calcul, nous avons jugé bon de générer un graphe qui contient uniquement les pires des temps de calculs observés. Le graphe obtenu est présenté dans la figure suivante :

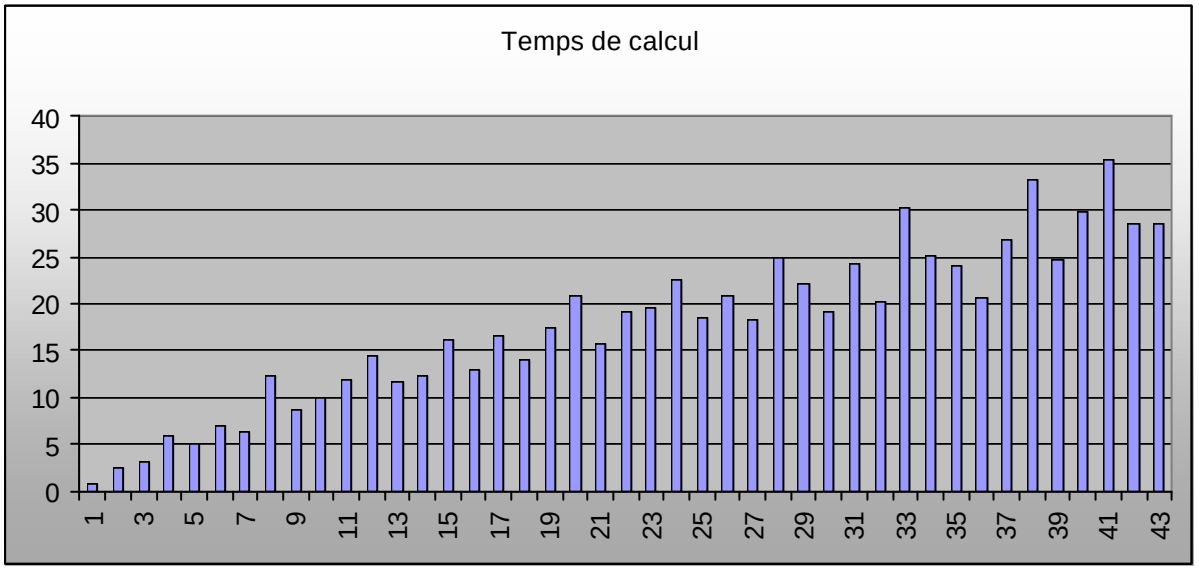


Figure 85. les temps de calculs sur les pires des cas observés par longueur

Notre observation sur la totalité des énoncés s'est confirmée avec l'analyse des pires des cas qui montre globalement un comportement linéaire comme nous pouvons le voir dans la figure précédente. Les exceptions à l'augmentation des temps de calculs progressivement avec l'augmentation des longueurs des énoncés sont dues à la différence en terme de fréquence entre les différentes longueurs.

La raison principale pour laquelle la complexité observée est linéaire est que la grammaire écrite ne contient pas des règles pour représenter des phénomènes qui causent un temps cubique pour les CFG

comme l'auto-enchâssement *center-self-embedding* qui se manifeste dans des énoncés du type : la souris qui craint le chat qui craint le chien qui est possédé par le voisin qui cherche un boulot. Ce genre d'énoncés peut être parfaitement traité dans le cadre du formalisme Sm-TAG mais la raison pour laquelle nous n'avons pas de règles pour le traiter dans notre grammaire est que nous ne l'avons pas observé dans notre corpus. En d'autres termes, les règles d'inférence et RTRs utilisées dans notre implantation (correspondantes aux arbres locaux et globaux) sont soit linéaires gauches soit linéaires droites et la grammaire réellement implantée est équivalente à une grammaire régulière.

2.1.1.84 Evaluation quantitative

Différents tests quantitatifs ont été réalisés sur Oasis dans les différentes étapes de son développement. Dans ce qui suit, nous allons nous limiter à la présentation du dernier test réalisé sur la dernière version du système afin d'éviter les confusions. Les lecteurs qui désirent en savoir plus sur ces tests intermédiaires peuvent consulter notre papier (Kurdi, 2000b).

2.2.4.1.1 Le corpus de test

Le corpus de test utilisé contient 210 énoncés transcrits extraits du corpus de réservation hôtelière. Les énoncés choisis font partie des dialogues non utilisés pour l'écriture de la grammaire. Afin de pouvoir tester le système sur la sortie de reconnaissance de la parole, nous avons procédé à une lecture de ces énoncés. Lors de la lecture des énoncés, nous avons simulé une prosodie spontanée afin d'obtenir des résultats proches des énoncés de dialogues réels. Après l'enregistrement, nous avons passé les fichiers son obtenus au système de reconnaissance RAPHAEL qui a donné comme sortie la meilleure hypothèse de reconnaissance correspondant à chacun des énoncés. La liste des sorties du systèpme de reconnaissance est donnée dans l'annexe de cette thèse.

2.2.4.1.2 Les résultats de l'évaluation

Deux unités ont été retenues pour le calcul des résultats :

1. **Le mot :** nous avons calculé le pourcentage des mots analysés sans prendre en considération le fait que l'analyse soit correcte ou pas. En d'autres termes, nous avons calculé le rapport du nombre des mots analysés *vs*. nombre des mots non analysés. La raison principale d'utiliser cette approche purement quantitative ici est de donner une idée sur la couverture lexicale du système et donc la portée de la stratégie sélective.
2. **L'arbre :** dans les statistiques nous avons considéré les arbres globaux et les arbres locaux. Contrairement au pourcentage des mots, ce critère est basé sur une distinction qualitative de l'analyse des unités. Ainsi, nous avons distingué entre trois types d'erreurs d'analyse des arbres :
    i. **Insertion :** lorsque le système ne supprime pas un élément qui doit être supprimé (élément répété, un relatif dont la complétive est supprimée, etc.) alors cet élément est considéré comme inséré.

ii. **Non-analyse :** seuls les arbres **pertinents** pour la tâche qui ne sont pas analysés (qu'ils soient couverts par notre corpus de base ou pas) sont considérés comme des cas de non-analyse. Ainsi, lorsque le système ignore à l'aide la stratégie sélective un arbre non pertinent nous ne considérons pas cela comme un cas de non-analyse.

iii. **Analyse incorrecte :** il s'agit des cas où le système associe une mauvaise analyse à un arbre pertinent pour la tâche.

Les résultats obtenus sur l'analyse de la meilleure hypothèse de reconnaissance de RAPHAEL sont présentés dans le tableau suivant :

Pourcentage des mots analysés	Rappel %	Précision %
66,24	83,72	96,77

Tableau 15. Résultats de l'évaluation sur la sortie de reconnaissance

Comme nous pouvons le remarquer dans le tableau précédent, le rappel de notre système est acceptable : environ 84% des arbres de notre corpus de test sont correctement analysés. Par ailleurs nous pouvons noter le taux de précision qui est assez élevé : 96,77%. Les raisons des erreurs que nous avons observées sont réparties sur les trois points suivant :

- Erreurs de reconnaissance : 53,58% des erreurs d'analyse sont causées par des erreurs de reconnaissance de différents types. Cependant, les erreurs de reconnaissance n'étaient pas une cause systématique d'erreurs d'analyse. En effet, nous avons remarqué que dans 34,78% des cas le système réussissait à donner une analyse correcte malgré l'existence d'une erreur de reconnaissance.
- Sous-génération de la grammaire : la sous-génération de la grammaire a été la cause de l'erreur d'analyse dans 42,85% des cas.
- Cas complexes linguistiquement : les cas linguistiquement complexes comme des ellipses spéciales, des incises, des anaphores, etc. ont causé des erreurs dans 3,57% des cas.

2.2.4.1.3 Comparaisons avec d'autres travaux

Les résultats du système de transport public (basé sur le formalisme de HPSG) préparé dans le cadre du projet hollandais OVIS (Nederhof *et al.*, 1997) présente un rappel de 87,4% et une précision de 85,5%. Par ailleurs, le système L'ATIS du LIMSI (Minker et Bennacef, 1996), (basé sur une grammaire sémantique de cas), donne un taux de 81,8% de réponses correctes (les auteurs n'ont pas donné la précision). Comparés à celles de notre système[46], ces résultats nous permettent de constater la

[46] Il faut noter que la comparaison avec les résultats des autres systèmes est approximative. D'une part, leurs corpus, leurs tâches de dialogue et la sortie de leur système (représentation sémantique ou arbre d'analyse

bonne performance de notre système en terme de précision et une performance acceptable en terme de rappel (même si la comparaison ne favorise pas notre système puisque les sorties du module sémantique utilisées pour tester les systèmes OVIS et L'ATIS tendent à avoir un rappel plus élevé que les systèmes qui ont pour sortie un arbre d'analyse syntaxique comme le nôtre).

2.1.1.85 Evaluation qualitative : la campagne d'évaluation par défi

Etant donné que l'objectif principal de notre réalisation du système Oasis est d'analyser la portée et la limite du formalisme Sm-TAG ainsi que celles du cadre de traitement des extragrammaticalités que nous avons proposé, il nous semble utile d'effectuer une évaluation qualitative de ce système en terme de couverture des phénomènes linguistiques. Pour ce faire, il nous faut des corpus de test appropriés. En effet, l'un des principaux obstacles devant ce genre d'évaluations est la difficulté à trouver des ressources linguistiques dans lesquels les phénomènes sont suffisamment représentés. Ainsi, nous avons décidé d'adopter une approche qui permet d'obtenir ce genre de données et de les utiliser pour évaluer Oasis. Il s'agit de l'approche d'évaluation par défi qui est une version simplifiée de la méthode DCR (Antoine *et al.*, 2000). Le principe général de cette méthode consiste à générer (par un ensemble de sujets humains) un corpus de test avec le maximum possible de phénomènes linguistiques à partir d'un petit corpus représentatif de la tâche dit corpus initial. Ainsi, d'une part, grâce au corpus initial la génération d'énoncés non pertinents pour la tâche du système testé devient très limitée et d'autre part, cela permet d'avoir une représentativité significative des phénomènes linguistiques qui sont l'objectif de l'évaluation. Les principales propriétés et démarches de cette évaluation sont décrites dans les points suivants :

2.2.4.1.4 Cadre de l'évaluation

Cette évaluation a été menée dans le cadre d'une campagne du GT "Compréhension robuste de la langue" du GDR-I3. Cinq systèmes représentant quatre laboratoires français sont impliqués dans cette campagne (Antoine *et al.*, 2002) :

---

syntaxique) ne sont pas identiques aux nôtres et d'autre part, les trois systèmes ont été testés dans des conditions différentes et avec des méthodes différentes.

Laboratoire	Système	Domaine	Responsable(s)
CLIPS-IMAG	Oasis	Réservation hôtelière	M. Z. Kurdi
IRIT	Cacao	Informations ferroviaires	C. Bousquet-Vernhettes et N. Vigouroux
LIMSI	Arise	Informations ferroviaires	S. Rosset
VALORIA	Logus[47]	Informations touristiques	J. Villaneau
VALORIA	Romus[48]	Informations touristiques	J. Goulian

Tableau 16. Les laboratoires et les systèmes impliqués dans la campagne d'évaluation par défi

Comme nous pouvons le remarquer dans le tableau précédent, les différents systèmes impliqués ont des domaines d'applications assez différents. Par ailleurs, les approches et les types de sortie de ces systèmes sont assez hétérogènes eux aussi (pour plus de détails sur ces systèmes voir (Antoine *et al.*, 2002)). Ainsi, vue ces différentes hétérogénéités, l'objectif de cette évaluation, dans l'étape actuelle, n'est pas de comparer directement les systèmes à la manière des campagnes de test DARPA-ATIS (Minker et Bennacef, 1996). En effet, l'évaluation par défi vise essentiellement à donner une idée fine sur le comportement de chacun des systèmes impliqués dans la campagne en rapport avec l'approche dans le cadre de laquelle il s'inscrit.

2.2.4.1.5 Déroulement de la campagne d'évaluation par défi

Le déroulement de cette campagne a été fait selon les démarches suivantes :

2. **Création du corpus initial :** le corpus initial est composé de vingt énoncés que chacun des participants a proposés comme corpus de base. Il s'agit généralement d'énoncés extraits du corpus sur lequel le système est entraîné et que le système en question est capable de traiter correctement. La liste des énoncés initiaux que nous avons proposée comme corpus initial pour l'évaluation du système Oasis est présentée dans l'annexe 5.
3. **Création du corpus dérivé :** il s'agit de la modification structurale de chacun des corpus en reformulant les différents énoncés avec des constructions linguistiques différentes. Autrement dit, il s'agit de générer un ensemble d'énoncés similaires globalement à l'énoncé initial mais en y ajoutant un phénomène linguistique spécifique à chaque fois. Ces phénomènes ne sont pas définis a priori et ont été laissés au choix de chaque concepteur de test selon son expérience avec son système (le nom par défi vient du fait que chaque participant essaie de générer des phénomènes qui peuvent poser un problème aux autres systèmes). Les phénomènes générés peuvent être des phénomènes grammaticaux (comme les extractions, les incises, les ellipses,

[47] (Villaneau *et al.*, 2002).

[48] (Goulian *et al.*, 2002).

etc.), des phénomènes extragrammaticaux (répétitions, hésitations, etc.) ou des simulations de phénomènes artificiels comme les erreurs de reconnaissance (les sujets suppriment, remplacent ou ajoutent des mots de manière similaire à ce qu'un système de reconnaissance peut faire en cas d'erreur). Ainsi, pour chaque énoncé initial chaque participant a crée quinze énoncés dérivés. Autrement dit, pour chaque énoncé initial nous avons obtenu soixante énoncés dérivés et un total de mille deux-cent énoncés dans le corpus dérivé. Voici, à titre d'exemple, un énoncé initial ainsi qu'un ensemble d'énoncés dérivés qui y correspondent :

i. L'énoncé initial :

<1> bon dans ces conditions alors réservez moi une chambre sympa et calme surtout pour le 26 février prochain </1>

ii. Cinq énoncés dérivés générés par notre collègue C. Bousquet de l'IRIT :

<1.1> bon dans ces conditions alors réservez moi **ben** une chambre sympa et **euh** calme surtout pour le 26 février prochain </1.1>

<1.2> bon dans ces conditions alors réservez moi une chambre sympa et calme surtout pour le 26 février **pro euh** prochain </1.2>

<1.3> bon dans ces conditions alors réservez moi une chambre **euh une chambre** sympa et calme surtout pour le 26 février prochain </1.3>

<1.4> bon dans ces conditions alors réservez moi une chambre sympa et calme surtout pour le 25 **euh non c'est pas ça 26** février prochain </1.4>

<1.5> bon dans ces conditions alors réservez moi une chambre sympa et calme surtout pour le 25 **euh 26 février** prochain</1.5>

Pour donner une idée plus précise sur l'opération de dérivation, un segment plus large du corpus dérivé sur lequel nous avons testé notre système est présenté dans l'annexe 5.

4. **Validation du corpus dérivé :** la validation consiste en le jugement par le créateur du système de l'adaptation des énoncés dérivés proposés. Les énoncés jugés non-adaptés font l'objet d'une modification par le créateur de test. Les principales demandes qui ont été faites par les participants portent sur les erreurs d'orthographe ainsi que sur les cas d'énoncés jugés non-pertinents ou non-réalistes par rapport à la tâche du système.

5. **Evaluation du système :** chacun des systèmes est évalué par son concepteur selon des critères qu'il juge appropriés. Le processus d'évaluation consiste en l'analyse et la classification des erreurs de chacun des systèmes lors de l'analyse des résultats. Les critères d'évaluation n'ont pas été définis *a priori*. Ainsi, chacun des participants a choisi la méthode de test qui lui semble la plus appropriée par rapport à son approche.

2.2.4.1.6 Les résultats du système Oasis

Avant de présenter les résultats de notre système dans le cadre de la campagne d'évaluation par défi, voici les deux points qui distinguent notre évaluation de celle des autres systèmes impliqués dans cette campagne :

- **Corpus considéré :** étant donné que la taille du corpus de test a augmenté au cours de la campagne (le LIMSI s'est joint à la campagne après son démarrage) et étant donné que nous ne savons pas *a priori* la fréquence des phénomènes linguistiques dans les corpus de test, nous avons décidé de faire le test uniquement sur un sous-ensemble des énoncés obtenus. Cela nous permettra d'analyser finement cette partie et en cas de constatation du besoin de plus de données afin d'avoir plus de représentativité pour les phénomènes nous pouvons ajouter une autre partie du corpus. Ainsi, nous avons pris les huit premiers groupes d'énoncés de chaque participant (qui correspondent chacun à l'ensemble des énoncés dérivés d'un énoncé initial). Cela fait cent-vingt énoncés par concepteur de test et un total de quatre-cent quatre-vingt énoncés dérivés.
- **Méthode de calcul des résultats :** les résultats ont été calculés selon la même méthode utilisée pour le système Safir. En effet, nous avons distingué entre le mot et les arbres locaux et globaux dans nos calculs. De même, nous avons considéré trois types d'erreurs : insertion, non analyse et analyse incorrecte.

I. Résultats généraux : afin de donner une idée sur la différence de complexité des sous-corpus utilisés (chacun des sous-corpus correspond à l'ensemble des énoncés générés par un partenaire de la campagne), nous avons décidé de donner les résultats classés selon les sous-corpus. Les résultats de notre système à la fin de cette campagne sont donnés dans le tableau suivant :

Concepteur de sous-corpus	Pourcentage des mots analysés	Rappel des arbres analysés	Précision des arbres analysés
C. Bousquet (IRIT)	73,98	96,99	99,8
G. Goulian (VALORIA)	60,49	91,3	92,6
S. Rosset (LIMSI)	70,35	90,63	97,51
J. Villeaneau (VALORIA)	75,13	92,31	99,4
Total	69,98	92,80	97,32

Tableau 17. Résultats généraux du système Oasis dans la campagne d'évaluation par défi classés par type d'erreur et par concepteur de test

Pour montrer plus clairement le rapport entre la couverture des mots, d'une part, et le rappel et la précision des arbres d'autre part, nous avons jugé bon de présenter les résultats sous forme de graphe. Le graphe obtenu est présenté dans la figure suivante :

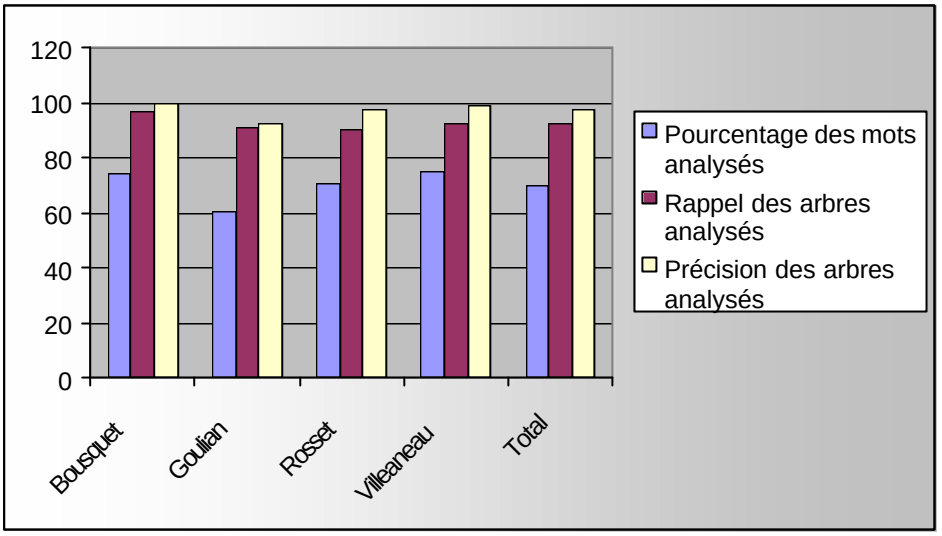


Figure 86. Les relations entre la couverture des mots, le rappel et la précision des arbres analysés par le système Oasis

Comme nous pouvons le remarquer dans le tableau 12**,** ainsi que dans la figure 89 la couverture lexicale n'a pas un rapport direct avec le rappel des arbres analysés. Cela montre à la fois l'efficacité de la stratégie sélective à localiser les îlots pertinents et l'absence d'effets négatifs de cette stratégie qui peuvent résulter de l'ignorance de segments couverts par la grammaire. Pour ailleurs, il n'est probablement pas inutile de signaler que les erreurs d'analyse présentées dans les tableaux ci-dessus ne correspondent pas forcément à des erreurs d'interprétation. En effet, après l'analyse des résultats, nous avons trouvé que 39,58% des erreurs d'analyse ne conduit pas à une erreur d'interprétation. Les deux principaux cas où une erreur d'analyse ne cause pas une erreur d'interprétation sont présentés dans les deux points suivants :

i. **Des cas d'insertion d'un arbre fonctionnel :** il s'agit généralement d'un pronom relatif ou une formule de demande qui ne peut pas être lié au reste de l'énoncé. Dans ce cas, n'importe quel module d'analyse sémantique raisonnablement robuste (IF, schéma, graphes conceptuels, etc.) devra exclure ces éléments non interprétables en dehors de leur contexte. Afin d'éclairer ce point, examinons l'exemple suivant :

Je voudrais une simple plutôt calme ***et dites-moi*** *si c'est avec douche ou bain*[49].

Après le traitement de l'énoncé précédent, le système produit l'analyse suivante :

[formule_demande

[49] Pour la facilité de l'exposé, cet exemple est une adaptation de l'énoncé (5.14 Goulian).

[je, pronom]
[voudrais, verbe]
]
[objet_demandé
[une, cardinal]
[simple, adjectif]
]
[coord
[conj, [plutôt, adverbe]]
]
[carc1
[calme, adjectif]
]
]
[coord
[et, conjonction]
]
[conj_cond
[si, pre]
]
[présentatif
[c, démonstratif]
[est, verbe]
]
[caractéristique_chambre
[avec, adverbe]
[une, cardinal]
[douche, nom]
]
[coord
[ou, conjonction]
]
[caractéristique_chambre
[bain, nom]
]

Dans cette analyse, nous remarquons que le système n'a pas associé une représentation quelconque au segment *dites-moi* à cause d'un problème de sous-génération de la grammaire. Cependant, à cause de la stratégie d'analyse partielle, le système a produit la conjonction de coordination *et* qui dépend du segment non analysé. L'analyse du mot *et* est considérée comme une erreur d'insertion mais d'un point de vue sémantique cela n'affecte pas l'interprétation.

ii. Des cas d'énoncés avec des éléments redondants : la redondance est l'une des sources principales de robustesse dans les langues naturelles en général (cela est valable à la fois

pour les systèmes naturels ou artificiels). Dans notre corpus de test, le cas que nous avons observé le plus fréquemment est la redondance de la formule de demande. Ainsi, dans un bon nombre d'énoncés il y a une formule de demande principale: *je voudrais, pouvez vous,* etc. couplée avec une formule auxiliaire souvent utilisée pour la politesse *s'il vous plaît, si c'est possible, etc.* ou pour la précision de la première formule *je préfère, de préférence,* etc. Dans ce genre d'énoncés, si pour une raison ou une autre le système n'arrive pas à analyser l'un des deux segments redondants cela est considéré comme une erreur d'analyse même si le système est capable à l'aide d'un des deux segments seulement de juger qu'il s'agit d'une requête.

II. **Analyse de l'effet de la complexité des énoncés sur la qualité d'analyse :** afin d'évaluer l'effet de la complexité des phénomènes linguistiques et artificiels sur la complétude, nous avons jugé bon de présenter les résultats du test selon les critères de complétude d'analyse. Cela donne, par ailleurs, une idée sur l'importance de la stratégie sélective dans le traitement (les énoncés qui ne sont pas entièrement traités dans les systèmes classiques sont rejetés).

Concepteur de sous-corpus	Pourcentage des énoncés dont tous les mots ont été analysés	Pourcentage des énoncés dont le rappel est de 100%	Pourcentage des énoncés dont la précision est de 100%
C. Bousquet (IRIT)	5,83	81,66	98,33
G. Goulian (VALORIA)	5	41,90	79,16
S. Rosset (LIMSI)	9	50,47	90
J. Villaneau (VALORIA)	4,16	50,83	95,84
Total	5,99	56,21	90,83

Tableau 18. Les résultats du système Oasis d'un point de vue complétude d'analyse

Comme nous pouvons le remarquer dans tableau précédent, le pourcentage des énoncés dont les mots sont entièrement analysés est très faible : 5,99. Cela veut dire, qu'en cas d'adoption d'une stratégie d'analyse classique, seule une partie mineure des énoncés du corpus aurait pu être analysée correctement. Par ailleurs, nous remarquons que les arbres pertinents dans un peu plus de la moitié des énoncés ont été analysés et que 90,83% des énoncés ont une précision de 100%. Cela montre que les erreurs de couverture (représentées par le rappel) sont plus réparties dans le corpus d'analyse que celles de précision.

III. **Analyse qualitative des résultats du système Oasis :** pour donner une idée des performances de notre système pour le traitement des principaux phénomènes linguistiques observés dans notre corpus de test, nous avons jugé bon de faire une analyse détaillée du traitement de ces

phénomènes. Afin de tenir en compte l'aspect sélectif de notre approche, nous avons considéré dans notre test toutes les occurrences d'un phénomène avec la distinction des phénomènes selon leur pertinence par rapport à la tâche. Ainsi, nous avons distingué entre deux types de cas :

- **Des cas valides :** il s'agit des cas qui se trouvent dans une zone pertinente par rapport à la tâche du système. Ces cas peuvent être positifs (correctement traités) ou négatifs (non traités ou incorrectement traités).
- **Des cas neutres :** ce sont des cas localisés dans la zone non pertinente de l'énoncé. Ces cas n'ont pas été considérés dans le calcul des pourcentages des phénomènes traités parce que notre objectif dans cette étape est d'analyser la performance du système en terme de couverture des phénomènes linguistiques pas la couverture lexicale.

  Pour la facilité de la présentation, nous avons distingué entre deux groupes de phénomènes : les phénomènes extragrammaticaux (les extragrammaticalités) et les phénomènes grammaticaux.

1. **Résultats du système Oasis pour le traitement des extragrammaticalités :** dans ce groupe nous avons distingué cinq phénomènes. Il s'agit des hésitations, mots incomplets, répétitions, autocorrections et faux-départs. Les résultats du système Oasis classés par sous-corpus sont présentés dans le tableau suivant :

Concepteur de sous-corpus	Nature des cas	Hésitation	Mot incomplet	Faux-départ	Autocorrection	Répétition	Total
C. Bousquet (IRIT)	Nombre total des cas	33	9	2	30	24	98
	Nombre des cas neutres	0	0	0	3	0	3
	Nombres des cas positifs	33	9	2	27	24	95
	Pourcentages des cas corrects	100	100	100	90	100	96,93
G. Goulian (VALORIA)	Total des cas	35	0	15	32	15	97
	Total des cas neutres	0	0	6	12	1	19
	Total des cas positifs	35	0	9	20	14	78
	Pourcentages des cas corrects	100	-	60	74,07	93,33	84,78
S. Rosset (LIMSI)	Nombre total des cas	21	2	7	17	13	60
	Nombre des cas neutres	0	0	3	4	3	10
	Nombres des cas positifs	21	2	4	13	10	50
	Pourcentages des cas corrects	100	100	57,14	76,47	76,92	83,33
J. Villaneau (VALORIA)	Total des cas	2	0	1	10	0	13
	Nombre des cas neutres	0	0	1	2	0	3
	Total des cas positifs	2	0	0	8	0	10
	Pourcentages des cas corrects	100	-	-	80	-	83.33
Total	Total des cas	91	11	25	89	52	268
	Nombre des cas neutres	0	0	12	21	4	37
	Total des cas positifs	91	11	13	68	48	231
	Pourcentages des cas corrects	100	100	62, 5	80, 95	92, 30	89,53

Tableau 19. Les résultats du système Oasis pour le traitement des extragrammaticalités classés par phénomène et par concepteur de sous-corpus

Comme nous pouvons le remarquer dans le tableau précédent, les résultats globaux obtenus sur les extragrammaticalités montrent que la performance du système Oasis pour le traitement des extragrammaticalités est proche de 90%. Cela peut être considéré comme une confirmation globale de nos résultats obtenus avec le système Corrector. Voici une discussion détaillée des résultats par phénomène :

i. **Les résultats sur les hésitations et les mots incomplets :** ces résultats montrent une efficacité pratiquement parfaite de notre approche pour le traitement de ces phénomènes. La raison principale de ce succès est la fonction de filtrage qui permet au système de filtrer tous les mots qu'il ne peut pas traiter.

ii. **Les résultats sur les répétitions et les autocorrections :** sur ce plan, les résultats sont globalement satisfaisants, en particulier vu l'état de complétude du module de traitement par patrons. Les cas d'échec observés sont principalement dus à la sous-génération et à des cas particulièrement difficiles (notamment à cause de zones d'édition compliquées). Comparé aux résultats obtenus avec le système Corrector, nous trouvons que les résultats obtenus avec Oasis sont légèrement supérieurs pour les répétitions alors qu'elles sont inférieures d'environ 5% pour l'autocorrection. L'explication de ces résultats est difficile à faire. En effet, un nombre assez important de variables distingue les deux évaluations comme la langue (la morphologie de l'anglais est moins riche que celle du français), la complexité des énoncés de test (les énoncés du *TRAINS corpus* nous semble plus complexes en termes d'extragrammaticalités que ceux du corpus collecté pour l'évaluation par défi).

iii. **Les résultats sur les faux départs et les incomplétudes :** nous remarquons que les résultats baissent comparés aux phénomènes précédents mais restent assez proches de ceux obtenus avec le système Oasis pour ces phénomènes respectifs. Il reste à dire que les erreurs de traitement des extragrammaticalités, comme les autres phénomènes que nous avons vus, ne mènent pas automatiquement à une erreur d'interprétation. Par ailleurs, il n'est cependant pas inutile de noter que la stratégie sélective a joué un rôle clé dans le traitement de ces phénomènes étant donné que le module de post-traitement est loin d'être complet. Cela montre que l'approche collaborative que nous avons adopté (collaboration de la stratégie sélective et des règles de post-traitement pour la détection et la délimitation) des faux-départs et des incomplétudes est prometteuse.

2. **Résultats du système Oasis pour le traitement des phénomènes grammaticaux :** nous avons distingué dans ce group cinq phénomènes. Il s'agit des ellipses, incises, extractions,

anaphores, négations, coordinations, ambiguïtés, erreurs de reconnaissance et ambiguïtés. Les résultats du système sur ces phénomènes sont présentés dans les tableaux suivants :

Concepteur de sous-corpus	Classification des cas selon leur nature	Ellipse	Incise	Extraction	Anaphore	Négation	Coordination	Ambiguïté	Erreurs de RAP	Relative	Total
C. Bousquet (IRIT)	Nombre total des cas	12	9	15	3	7	17	0	33	192	288
	Nombre des cas neutres	0	0	0	0	0	0	0		0	0
	Nombres des cas positifs	12	9	15	3	7	17	0	15	171	249
	Pourcentages des cas corrects	100	100	100	100	100	100	-	45,45	94,13	86,45
G. Goulian (VALORIA)	Total des cas	35	54	71	32	33	21	2	0	26	274
	Total des cas neutres	3	8	2	15	18	6	0	0	24	76
	Total des cas positifs	30	39	68	15	12	12	1	0	2	179
	Pourcentages des cas corrects	93,75	84,78	98,55	88,23	80	80	50	-	100	90,40
S. Rosset (LIMSI)	Nombre total des cas	31	20	41	6	2	46	0	0	9	155
	Nombre des cas neutres	5	0	0	4	2	4	0	0	4	17
	Nombres des cas positifs	26	20	41	2		42	0	0	5	136
	Pourcentages des cas corrects	100	100	100	100	-	91,30	-	-	100	98,55
J. Villaneau (VALORIA)	Total des cas	20	14	72	37	11	48	0	0	10	212
	Nombre des cas neutres	1	0	5	15	1	4	0	0	4	30
	Total des cas positifs	18	12	63	18	10	40	0	0	2	163
	Pourcentages des cas corrects	100	85,71	94,02	81,81	100	100	-	-	100	89,56

Tableau 20. Résultats du système Oasis pour le traitement des phénomènes grammaticaux classés par phénomène et par concepteur de sous-corpus

Classification des cas selon leur nature	Ellipse	Incise	Extraction	Anaphore	Négation	Coordination	Ambiguïtés	Erreurs de RAP	Relative	Total
Nombre total de tous les cas	98	97	199	78	53	132	2	33	237	929
Nombre total de tous les cas neutres	9	8	7	34	21	14	0	0	32	123
Nombre total de tous les cas positifs	86	80	187	38	29	111	1	15	205	717
Pourcentages totaux des cas corrects	96,62	89,88	97,39	86,36	90,62	94,06	50	45,45	87,8	88,95

Tableau 21. Résultats globaux du système Oasis pour le traitement des phénomènes grammaticaux

Les tableaux précédents montrent que la performance globale de notre système est assez bonne pour le traitement des phénomènes linguistiques observés dans notre corpus de test : 88,95% des cas ont été correctement analysés. Dans ce qui suit, nous allons faire une analyse détaillée des résultats de chaque phénomène à part.

i. **Le traitement des ellipses :** la plupart des ellipses observées dans notre corpus de test sont des ellipses verbales (omission du verbe ou d'une construction verbale). Des ellipses d'autres éléments sont aussi observées comme celles du déterminant d'un nom. Nous avons eu, au total, quatre-vingt-neuf cas valides d'ellipses dont quatre-vingt-six ont été correctement traités (96,62%). Globalement, les cas qui n'ont pas été traités ne correspondent pas à des formes courantes d'ellipse (nous n'avons pas observé des cas similaires dans le corpus de réservation hôtelière). Par exemple, nous avons eu des cas difficiles d'ellipse du déterminant d'un nom qui ont causé une erreur de traitement comme dans :

Train arrive 10 12 19 heures 37 <2.11, Goulian> (108)

Ces phénomènes peu fréquents ne causaient pas une erreur systématique, c'est-à-dire, le système a été capable de traiter des suppressions de déterminants de noms comme dans l'exemple :

8 octobre une baignoire si c'est possible (109)

Par ailleurs, le système a très bien réussi à traiter les ellipses verbales qui sont assez courantes dans le dialogue comme l'ellipse de la construction verbale *je voudrais* de l'énoncé précédent.

ii. **Le traitement des incises :** l'incise consiste à insérer un mot, un segment ou un énoncé entier entre deux unités qui sont généralement connectées l'une à l'autre et dont la connexion est nécessaire pour juger la grammaticalité de ces deux unités. Qutre-vingt-neuf cas valides ont été observés dans notre corpus de test dont quatre-vingt ont été correctement traités (environ 90%). Les erreurs de traitement sont dues à l'insertion de segments non pertinents : comme l'expression couramment utilisée dans les incises *je veux dire* qui n'est pas modélisée dans notre grammaire. Le système, dans ce cas, considère *dire* comme un mot non pertinent et insère *je veux* comme une formule de demande. Ces erreurs comme la plupart des erreurs d'insertion ne posent pas un problème pour l'interprétation de l'énoncé étant donné que le segment inséré est non pertinent par rapport au contexte.

iii. **Le traitement des anaphores :** dans notre corpus de test, trente-huit cas valides d'anaphores ont été observés. Trente-quatre cas ont été correctement traités, c'est-à-dire, 86,36% ont été correctement traités. Le pourcentage assez élevé des cas neutres est dû aux reprises anaphoriques fréquentes de segments non pertinents ou d'insertion de verbes de l'extérieur du lexique. La

partie principale des erreurs est due à la sous-génération de la grammaire. En effet, un bon nombre des constructions non traitées n'a pas été observé lors de l'écriture de la grammaire. Dans ces cas, le système considère le pronom anaphorique comme un mot inséré et échoue à traiter tout le segment.

iv. **Le traitement des extractions :** les extractions consistent à déplacer un segment d'un endroit à un autre dans l'énoncé. Les segments déplacés sont généralement des syntagmes prépositionnels dont la position a été changée pour mettre l'accent sur leur contenu. Nous avons observé cent quatre-vingt-deux cas valides dans notre corpus de test. Un bon pourcentage de ces cas a été correctement traité: 97,39%. Cela est dû notamment grâce à la stratégie d'analyse partielle qui permet à des unités non connectées avec le reste de l'énoncé (comme c'est le cas des unités déplacées dans l'extraction) d'être considérées comme des unités bien formées. Les erreurs de traitement sont dues à des extractions d'unités inférieures à un arbre local.

v. **Le traitement des négations :** la négation est un phénomène syntaxique assez important dans la mesure où il est directement impliqué dans l'interprétation de l'énoncé. Dans notre corpus de test, nous avons observé trente-deux cas valides de négation dont vingt neuf ont été correctement traités, c'est-à-dire 90,62% des cas. Les trois erreurs observées sont dues à la sous-génération de la grammaire.

vi. **Le traitement des coordinations :** nous avons observé cent dix-huit cas valides de coordination dans le corpus de test. Cent onze ont été correctement traités par le système c'est-à-dire 94,06%. Au cours de notre analyse nous avons remarqué un bon traitement de toutes les formes d'extraction qui ont impliqué des arbres locaux notamment grâce à l'approche d'analyse partielle. La raison principale de l'échec est les coordinations d'éléments au sein même d'un arbre local.

vii. **Traitement des ambiguïtés :** par ambiguïtés, nous entendons tous les cas d'ambiguïtés syntaxiques qui ne sont pas couverts par les autres phénomènes considérés dans notre classification comme les problèmes de portée de la négation ou de la coordination qui sont traités avec leurs phénomènes respectifs. Un des deux cas d'ambiguïté n'a pas été correctement traité. Il s'agit d'un cas difficile de rattachement de *syntagme prépositionnel* post-posé sans reprise anaphorique.

viii. **Traitement des erreurs de RAP:** seule C. Bousquet de l'IRIT a produit des énoncés qui contiennent des simulations d'erreurs de reconnaissance. Elle a produit trente-trois erreurs de différents types (insertion, suppression, remplacement). Dans quinze cas (c'est-à-dire 45,45% des cas), le système a réussi à rattraper ces erreurs. Le rattrapage a été réalisé dans les cas ou les erreurs de reconnaissances ont endommagé une partie non centrale dans le traitement.

ix. **Traitement des relatives :** les constructions relatives sont le phénomène le plus fréquent que nous avons observé dans notre corpus de test avec un nombre total de deux cent-cinq cas valides. Environ 88% des cas d'énoncé avec une construction relative ont été correctement traités. La raison principale des erreurs d'analyse est la sous génération de la grammaire. En effet, la majorité des échecs est due à des expressions composées d'un relatif et un verbe inconnu (non couvert dans le dictionnaire du système) comme *qui vient, dire que*, etc.

Comme nous pouvons le constater à travers notre discussion des résultats des phénomènes grammaticaux et extragrammaticaux, le taux de bon traitement du système Oasis est généralement assez élevé. Nous avons vu aussi que les raisons principales d'échec de l'analyse sont liées essentiellement à l'état actuel du système en terme de développement ou à la disponibilité des données qui est la source majeure des problèmes de sous-génération. Ainsi, nous pouvons conclure que notre approche (basée sur la Sm-TAG, traitée par un algorithme d'analyse partielle et sélective, couplée avec l'approche de traitement des extragrammaticalités) combine raisonnablement bien la robustesse et la profondeur d'analyse.

2.2.4.1.7 Les premiers résultats globaux des systèmes impliqués dans la campagne

Les premiers résultats obtenus par les différents partenaires ont été présentés selon une typologie générale moins riche que celle que nous avons adoptée pour présenter les résultats de notre système dans les paragraphes précédents. Les motivations principales de cette simplification de la typologie sont la facilité de synthèse des résultats obtenus avec les quatre systèmes impliqués dans la campagne ainsi que des contraintes liées à certains partenaires. Les six phénomènes distingués dans cette typologie sont présentés dans les points suivants :

- Erreurs de reconnaissance de la parole : qui portent sur des cas d'insertion, suppression et remplacement de mots.
- Complexité structurale du langage oral : il s'agit des phénomènes syntaxiques et sémantiques complexes comme les coordinations, les négations, les subordonnées, etc.
- Les extragrammaticalité du langage oral : cela couvre les différents types d'extragrammaticalités lexicales et supralexicales.
- Les variations de l'ordre des mots : cela couvre les différentes formes de changement de l'ordre des mots dans l'énoncé comme : les extractions, les clivées, les interrogations par inversement de l'ordre des mots, etc.
- Couverture lexicale et sémantique : cela porte tant sur les mots pertinents non couverts par le lexique du système que sur les expressions non considérées dans le modèle sémantique (pour les systèmes qui comportent un module d'analyse sémantique).

- Phénomènes divers : il s'agit de phénomènes qui ne sont pas couverts par la typologie et qui sont d'intérêt particulier pour l'un des systèmes.

La méthode d'analyse des résultats qui a été retenue par les partenaires de la campagne consiste à calculer le pourcentage des cas où un phénomène n'a pas été traité correctement par rapport à la totalité des erreurs d'analyse. Par exemple, si nous avons 100 cas d'erreurs d'analyse au total et si 10 de ces cas sont causés par des ellipses, alors le pourcentage des erreurs causées par l'ellipse est de 10%.

Les résultats globaux des différents systèmes impliqués dans la campagne sont présentés dans le tableau suivant (Antoine *et al.*, 2002) :

Système / Type d'erreur	Oasis (CLIPS)	Cacao (IRIT)	Arise (LIMSI)	Romus (VALORIA)	Logus (VALORIA)
Erreurs de reconnaissance de la parole	7,0 %	0%	0%	20%	2%
Complexité structurale	12,5 %	2,8%	0%	6%	8%
Extragrammaticalités	9,0 %	6%	18,2 %	17%	32%
Variations de l'ordre des mots	2,3 %	14,9%	9,0 %	6%	3%
Couverture lexicale et sémantique	69,2 %	72,6%	36,0 %	32%	35%
Autres	-	3,7%	36,8 %	19%	20%

Tableau 22. Les résultats généraux des systèmes impliqués dans la campagne d'évaluation par défi

A première vue, nous pouvons remarquer que les résultats des différents systèmes sont hétérogènes (chacun des systèmes a un type d'erreur particulier). Malheureusement, il nous est impossible d'aller loin dans l'interprétation de ces résultats en les liant aux approches des systèmes. En effet, vu les différences entre les tâches des différents systèmes, les pourcentages des phénomènes linguistiques ainsi que leurs complexités ne sont pas identiques dans les différents corpus de test. Ainsi, nous ne pouvons pas distinguer si un pourcentage d'erreur peu élevé pour un phénomène quelconque peut être interprété comme une bonne performance du système ou comme une conséquence d'une fréquence peu élevée de ce phénomène dans le corpus de test.

Pour clarifier les résultats collectifs et rendre les comparaisons plus faciles, une évaluation plus avancée est en cours. Dans cette évaluation, une méthode de calcul similaire à celle que nous avons adoptée pour l'obtention des résultats de notre système sera adoptée : calcul du pourcentage des occurrences correctement traitées d'un phénomène sur la totalité des occurrences de ce phénomène. Par ailleurs, nous avons proposé une nouvelle méthode d'évaluation DCR étendue qui permet de générer objectivement les énoncés dérivés sur la base d'une grammaire générale (Kurdi et Ahafhaf, 2002). Cette génération objective rend la comparaison des résultats de différents systèmes plus faciles à faire étant donné que les énoncés produits ont le même degré de complexité.

Les premières expériences effectuées sur cette méthode avec le système Oasis ont montré qu'elle est prometteuse pour des applications larges similaires à celle de la campagne d'évaluation par défi. Une description de ces premières expériences ainsi que des résultats du système Oasis sont présentés dans l'annexe 6.

3 Chapitre III.3 : Le système Navigator pour la compréhension des dialogues mutli-domaines orientés par la tâche

Avec les développements dans le domaine de l'informatique et des télécommunications on assiste à une extension des domaines de dialogue. Ainsi, nous passons des dialogues mono-domaines orientés par la tâche aux dialogues orientés par la tâche et dont la tâche couvre plusieurs domaines. L'élargissement des domaines de dialogue implique l'élargissement du nombre des items lexicaux à considérer, l'augmentation des connaissances sémantiques et pragmatiques ainsi que les connaissances sur le domaine que le système doit prendre en considération lors du traitement. Comme la Sm-TAG intègre directement des connaissances sur le domaine, on pourrait penser que l'élargissement du domaine de dialogue peut avoir un effet sur les systèmes à base de Sm-TAG plus que les systèmes à base de formalismes syntaxiques classiques.

Dans ce chapitre nous allons présenter le système Navigator qui est une implantation de la Sm-TAG dans le contexte d'un dialogue multi-domaine. La propriété principale de Navigator est l'adoption d'une architecture hautement modulaire qui permet de réduire au maximum les inconvénients de la prise en considération des connaissances sur le monde au sein de la Sm-TAG. Ainsi, dans notre discussion et évaluation nous allons nous concentrer principalement sur les problèmes liés à l'élargissement du domaine de dialogue et leur effet potentiel sur la Sm-TAG tout en abordant les autres aspects du système pour donner une idée générale sur ses différentes composantes.

3.1 Le Projet Vico

Navigator a été réalisé dans le cadre du projet européen Vico qui a commencé au mois de Mars 2001 et dont la durée est de trois ans. Ce projet vise la construction d'un système de dialogue qui sert à contrôler un ensemble d'utilitaires dans la voiture comme le système de Guidage Par Satellite GPS ou l'accès à des informations générales via un réseau spécialisé appelé CWW (Car Wide Web). Les langues retenues pour ce projet sont l'anglais, l'allemand et l'italien. Cinq partenaires académiques et industriels sont impliqués dans ce projet. L'identité des partenaires ainsi que leurs contributions au projet sont présentés dans le tableau suivant :

LABORATOIRE	TYPE	LOCATION GEOGRAPHIQUE	PARTICIPATION AU PROJET
Bosch	Laboratoire industriel	Stuttgart, Allemagne	Coordination du projet et intégration des modules
Daimler Chrysler	Laboratoire industriel	Stuttgart, Allemagne	Reconnaissance de la parole pour l'anglais et l'allemand
IRST	Centre de recherche	Trento, Italie	Reconnaissance de la parole pour l'italien et Car Wide Web (CWW)
NISlab.	Laboratoire universitaire	Odense, Danemark	• Modules de compréhension pour les trois langues du projet. • Le gestionnaire de dialogue et les modules associés comme le profile d'utilisateur et le gestionnaire de la tâche. • Trois modules de génération pour les trois langues du projet.
Tele-Atlas	Laboratoire industriel	Belgique	Base de données géographiques

Tableau 23. Les partenaires du projet Vico et leur participation

Les principaux *challenges* de ce projet sont les suivants :

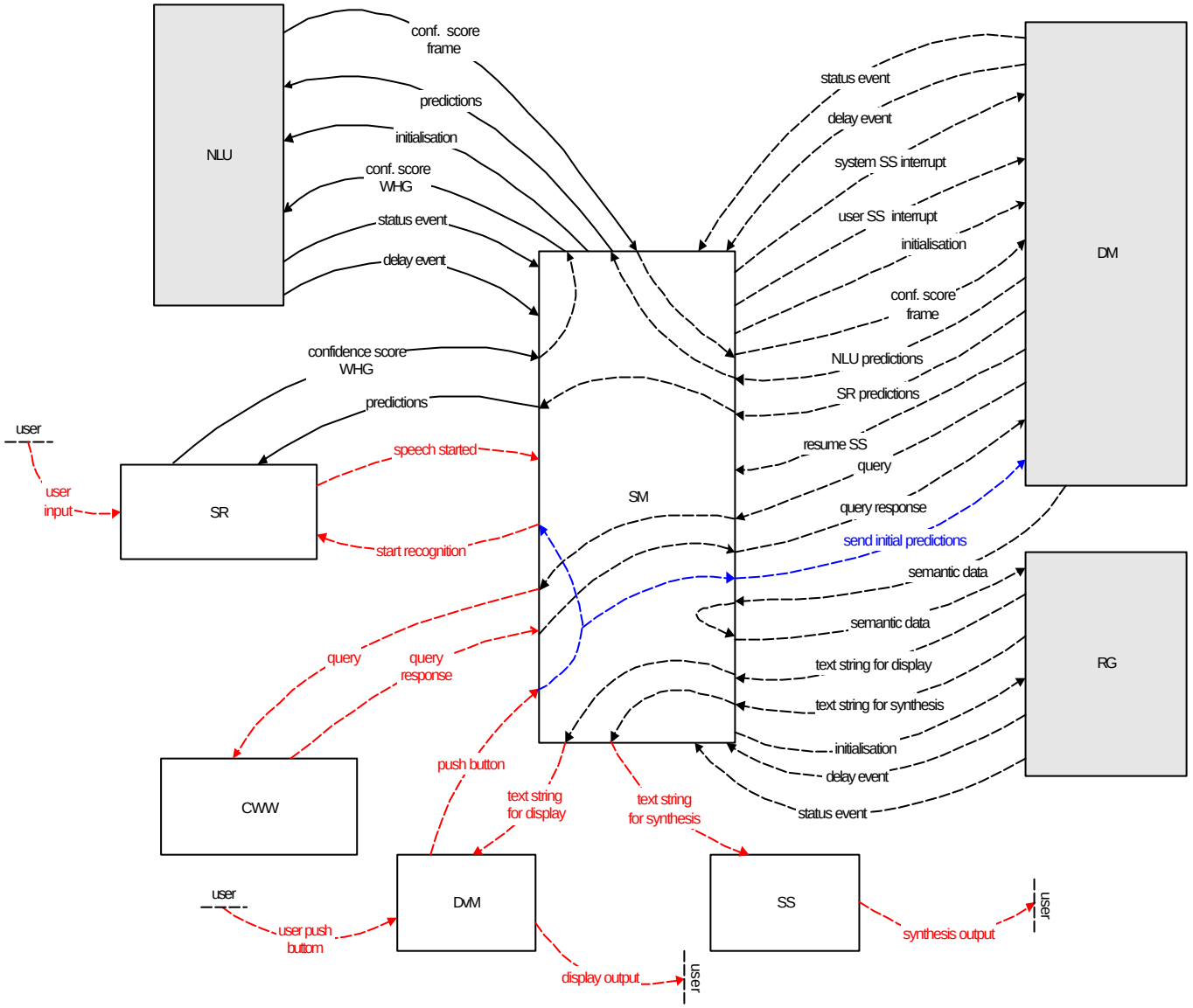
1. Pour le module de compréhension : comme nous avons dit, la largeur du domaine constitue le challenge principal pour le module de compréhension dans Vico.
2. Pour la reconnaissance: le nombre considérable des noms propres correspondants aux rues, villes, pays, points d'intérêts est d'environ 80.000 mots (dans la première étape du projet, près de 16000 mots sont utilisés). Par ailleurs, le bruit dans la voiture (à la fois le bruit du moteur et le bruit causé par les autres personnes et animaux dans la voiture) constituent aussi un challenge important à résoudre au niveau du traitement du signal de la parole.
3. Le gestionnaire du dialogue : vu le nombre potentiel d'erreurs de reconnaissance et d'ambiguïtés, on s'attend à ce que le déroulement du dialogue soit particulièrement difficile. Cela implique la mise en œuvre d'une approche de dialogue particulièrement souple et adaptative et qui permet de résoudre les

ambiguïtés en guidant le système de reconnaissance (en indiquant la région dans laquelle s'effectue la recherche et les thèmes possibles qui peuvent être abordés par l'utilisateur sachant le contexte dialogique) et aussi en déclenchant en cas de besoin des sous-dialogues de clarification.

3.2 Architecture du système Vico

L'architecture de Vico est basée sur un hub (gestionnaire de système) autour duquel sont organisés les différents modules. Deux propriétés clés de cette architecture méritent d'être cités :

1. Vico est un système hautement interactif : nous pouvons noter en particulier le rôle du gestionnaire de dialogue qui interagit avec la majorité des modules du système en fournissant des attentes et des instructions aux quatre modules de traitement linguistique : la reconnaissance, la compréhension, la génération et la synthèse et en effectuant des requêtes au CWW qui est la source principale des informations sur le monde dans le système.
2. L'interaction des modules se fait via CORBA (**C**ommon **O**bject **R**equest **B**roker **A**rchitecture) qui a été choisi en particulier à cause de l'hétérogénéité des modules et la nécessité d'accéder à des informations via le réseau avec le CWW.



NLU = Natural Language Understanding (Navigator) DM = Dialogue Manager SR = Speech recognition SM = System Manager CWW = Car Wide Web	RG = Response Generation SS = Speech Synthesis DvM = Device Manager

Figure 87. Architecture générale du système Vico (Bernsen, 2002)

3.2.1 Les modules de reconnaissance

Les modules de reconnaissance ont été développés Daimler Chrysler (pour l'anglais et l'allemand) et par les partenaires de l'IRST (pour l'italien). La spécificité principale des modules de reconnaissances utilisés dans Vico est leur modularité. En effet, afin de réduire les problèmes liés à la reconnaissance des mots géographiques (dont le nombre est assez élevé et qui ne peuvent pas être distingués facilement avec les modèles de langage puisque tous ces mots peuvent apparaître dans des contextes linguistiques similaires), les partenaires ont divisé le système de reconnaissance en unités de reconnaissance qui couvrent chacune un aspect particulier de la tâche. Ainsi, lorsqu'un signal de parole est détecté par le système de reconnaissance seul un sous-ensemble de ces unités est activé (le choix des unités à activer est basé sur les attentes du gestionnaire de dialogue). Les sept unités de reconnaissance utilisées ainsi que les techniques sous-jacentes à ces unités (grammaire ou Modèle Statistique de Langage) sont présentés dans le tableau suivant :

Unité	Technique utilisée	Domaine
SRU0	MSL	Navigation à Trentino
SRU1	MSL	Navigation à Bolzano
SRU2	MSL	Réservation hôtelière
SRU3	Grammaire	Méta-communications
SRU4	Grammaire	Les noms des villes et places dans la région du Trentino
SRU5	Grammaire	La liste des rues dans la région du Trentino
SRU6	Grammaire	Epellations

Tableau 24. Les unités de reconnaissance utilisées dans Vico

La sortie des systèmes de reconnaissance est une liste de N graphes où N est le nombre des unités de reconnaissance actives. Chaque graphe correspond à la meilleure hypothèse de l'unité active et contient la liste des mots reconnus couplés avec les scores individuels de reconnaissance pour chaque mot. Par ailleurs, chaque graphe est étiqueté par une catégorie qui représente l'unité de reconnaissance qui l'a fourni. Cela permettra au module de compréhension (qui prend ce graphe comme entrée) d'inférer l'attente du gestionnaire de dialogue associée à ce graphe. Ainsi, la sortie des systèmes de reconnaissance a le format suivant :

$SRU_1\ M_1\ Srm_1\ M_2\ Srm_2 \ldots M_a\ Srm_a \ldots SRU_n\ M_1\ Srm_1\ M_2\ Srm_2 \ldots M_c\ Srm_b$

Où :

1. SRU_x (Speech Recognition Unit) est l'indicateur de l'unité qui a produit l'hypothèse.
2. M_x est le mot reconnu (ou un modèle de bruit).
3. Srm_x est le score de reconnaissance associé à chaque mot.
4. *N* est le nombre des unités actives (et par conséquent celui des graphes fournis).
5. *a* et *c* correspondent respectivement aux longueurs des graphes 1et *N*.

Par exemple, pour un l'énoncé *I want to go to Trento* nous pouvons avoir la sorite suivante :

SRU_0 I 0.55 want 0.64 to 0.39 go 0.5 Trento 0.36 SRU_3 yes 0.31 #noise# 0.26 #noise# 0.28 SRU_4 #noise# 0.21 #noise# 0.18 Trentino 0.23

Comme nous pouvons le voir dans l'exemple précédent, trois unités de reconnaissance ont été activées. Il s'agit de l'unité de navigation routière dans la région du Trentino, l'unité des méta-communications et l'unité des noms de ville.

3.2.2 Le Gestionnaire de Dialogue (GD)

Ce module a été conçu et développé au NISLab principalement par nos collègues N. O. Bernsen, Laila Dybkjær et M. Charfuelan (Bernsen, 2002). Le GD est équipé de différentes fonctionnalités comme un gestionnaire de domaine (pour effectuer des raisonnements sur la cohérence des représentations sémantiques reçues), un modèle d'utilisateur (qui sert de mémoire à long terme du système), etc. En ce qui concerne le module de compréhension, deux propriétés nous semblent intéressantes à présenter avec plus de détails :

1. **Stratégie d'adaptation dynamique :** étant destiné à un dialogue dans des conditions assez variées (le niveau de bruit peut varier d'une voiture à une autre et dans la même voiture d'un moment à un autre selon les conditions naturelles comme la pluie ou autre), le gestionnaire de dialogue est équipé d'un mécanisme qui lui permet d'adopter la stratégie de dialogue la plus appropriées. Le choix de la stratégie est basé sur un score de confiance qui doit exprimer le degré de satisfaction de l'analyse sémantique reçue aux normes du module de reconnaissance de la parole et du module de compréhension.
2. **Production d'attentes :** le GD grâce à sa connaissance globale du contexte dialogique ainsi que du domaine de dialogue fournit des attentes qui guident les systèmes de reconnaissance et de compréhension. Ainsi, nous pouvons distinguer entre trois types d'attentes :
    i. Des attentes spécifiques à la reconnaissance : il s'agit des attentes relatives aux localisations géographiques des noms propres qui peuvent être abordés dans les énoncés de l'utilisateur. Comme le lexique du système de reconnaissance est organisé selon les zones géographiques, les attentes du GD permettent d'activer uniquement le lexique de la zone pertinente.

ii. Information fournie au module de compréhension sur la tâche courante : il s'agit de l'information sur le domaine du dialogue courant comme la navigation routière, le point d'intérêt, la réservation hôtelière, l'information sur Vico ou l'épellation. Lorsque le GD est incapable de fournir cette information (cela arrive au début du dialogue en général), il fournit l'étiquette *vide*.

iii. Des attentes communes : il s'agit de l'information fournie par le GD sur les domaines possibles qui peuvent être abordés par l'utilisateur sachant l'historique du dialogue. Les valeurs que peuvent prendre ces attentes sont identiques à celles des attentes du module de compréhension sauf que dans ce cas le GD fournit généralement plus d'un domaine.

3.3 Le module de compréhension[50] de Vico : Navigator

Ce système est conçu pour traiter des énoncés oraux en trois langues : l'anglais, l'allemand et l'italien. Le dialogue englobe : navigation routière, points d'intérêts, réservation hôtelière et information sur le système (aide). Les propriétés principales de Navigator peuvent être résumées dans les points suivants :

1. L'entrée du système est une liste de graphes de mots.
2. Adoption de la Sm-TAG comme formalisme d'analyse grammaticale.
3. Adoption d'une architecture modulaire à base de Hub. Le principe de base de cette architecture est la maximisation du partage des ressources linguistiques et logicielles d'une part à travers les trois langues et d'autre part à travers les différents domaines de dialogue.
4. Prise en considération des attentes du gestionnaire de dialogue dans la désambiguïsation des énoncés.

L'architecture générale de Navigator et ses interactions avec les autres modules de Vico sont présentées dans les deux figures suivantes :

[50] Navigator est un module de compréhension dans la mesure où il prend en considération le contexte dialgique dans ses analyes.

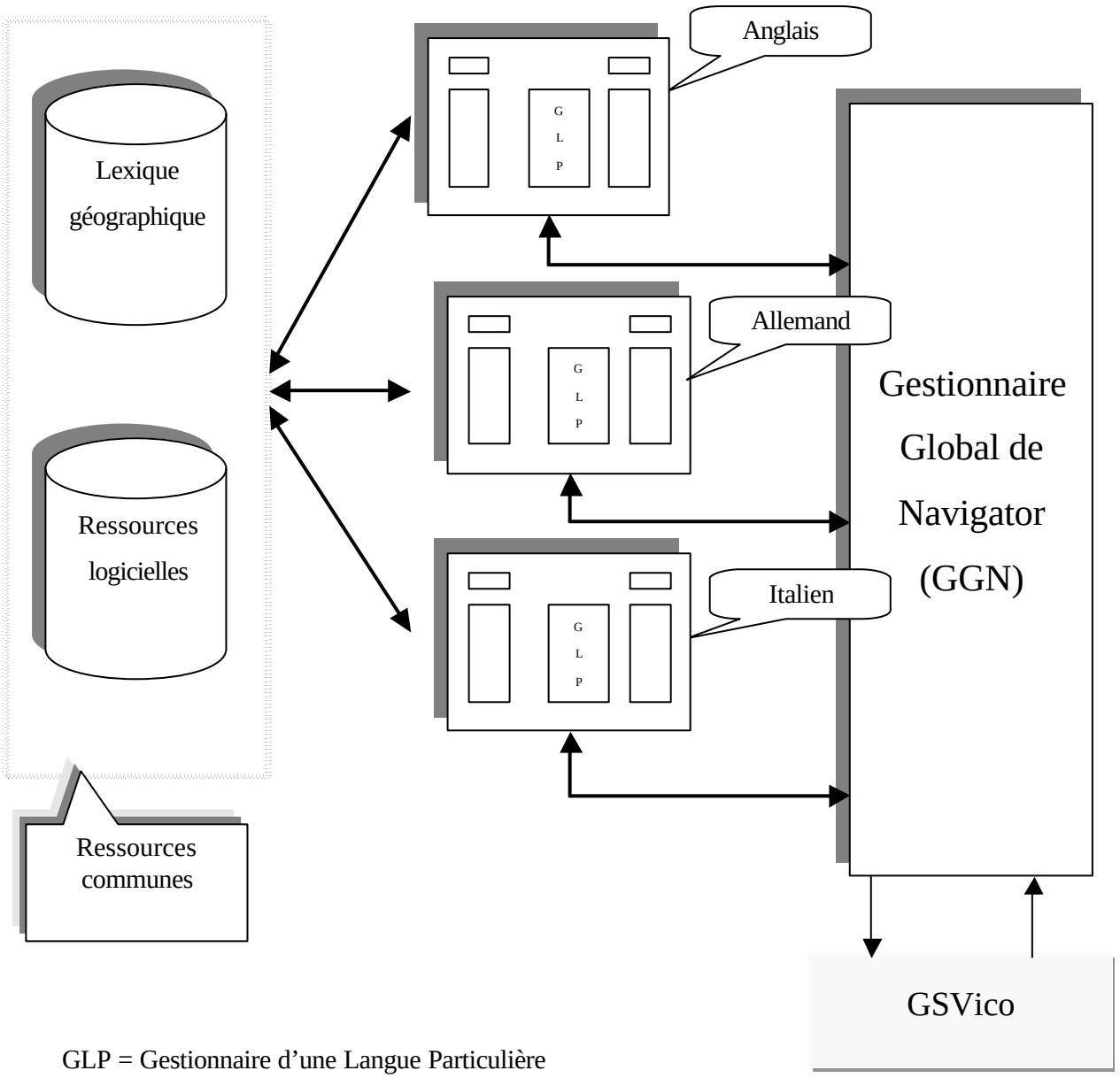
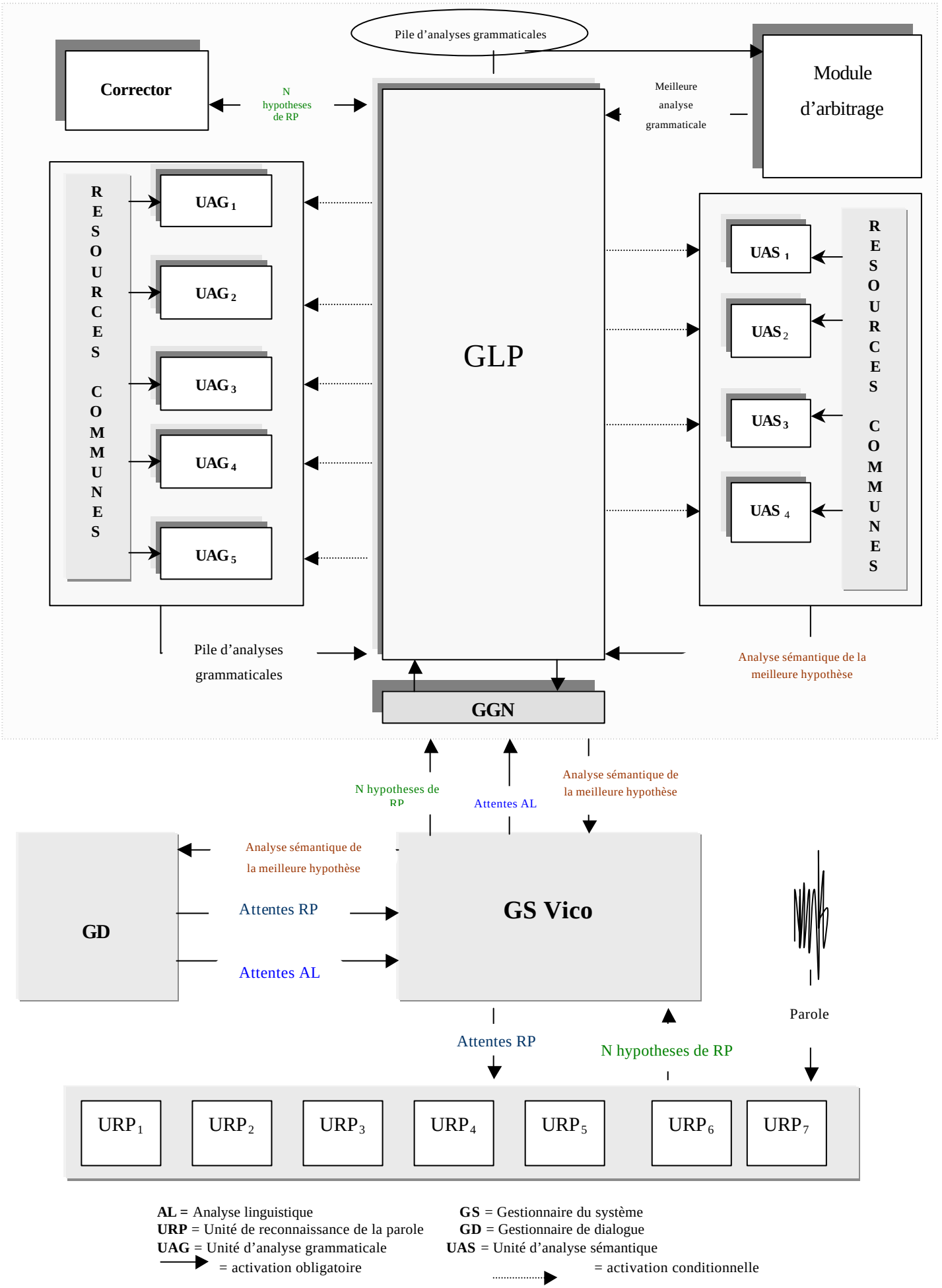


Figure 89. Architecture générale du module de compréhension Navigator



AL = Analyse linguistique
URP = Unité de reconnaissance de la parole
UAG = Unité d'analyse grammaticale
→ = activation obligatoire

GS = Gestionnaire du système
GD = Gestionnaire de dialogue
UAS = Unité d'analyse sémantique
⇢ = activation conditionnelle

Figure 90. Architecture de Navigator et ses interactions avec les autres modules de Vico

Le flux de l'information au sein du système Navigator est présenté dans le tableau suivant :

Entrée	Source	Destination	Traitement	Destination
N Meilleures hypothèse de reconnaissance	L'un des trois modules de reconnaissance (via le GS de Vico)	GGN	Activation de la langue appropriée selon l'attente du Gestionnaire de dialogue	GLP
N Meilleures hypothèse de reconnaissance	GLP	Corrector	Traitement des extragrammaticalités dans chacune des hypothèses reçues	GLP
N Meilleures hypothèse de reconnaissance dont les extragrammaticalités sont étiquetées	GLP	Analyse grammaticale : activation des unités d'analyse pertinentes	Association à chacune des hypothèses d'un arbre d'analyse Sm-TAG	GLP
N arbres d'analyse	GLP	Module d'arbitrage	Sélection la meilleure analyse grammaticale	GLP
Meilleure analyse grammaticale	GLP	Module d'analyse sémantique : activation d'une seule unité d'analyse sémantique	Schéma sémantique	GLP
Schéma sémantique	GLP	GGN	-	Le gestionnaire de dialogue (via le GS de Vico)

Tableau 25. Le flux de l'information au sein du système Navigator

Ainsi, nous pouvons dire que l'architecture de Navigator est à la fois une extension et une généralisation de celles de Corrector et d'Oasis.

3.3.1 Description des composantes de Navigator

3.3.1.1 Le Gestionnaire Global de Navigator (GGN)

Ce module sert d'interface entre d'une part le gestionnaire de Vico (et par conséquent le reste des modules du système de dialogue) et d'autre part les trois modules de compréhension correspondant aux trois langues du projet. Ainsi, selon l'information fournie par le gestionnaire de dialogue sur la langue courante, il active le module de compréhension approprié et lui envoie l'énoncé reçu. L'adoption d'une interface commune pour les trois langues est motivée par les deux raisons suivantes :

1. Elle facilite l'intégration du système : l'effort d'intégration pour les trois modules (pour les trois langues) est équivalent à celui d'un seul module.
2. Elle permet à tout moment du dialogue de passer d'une langue à une autre sans avoir à réinitialiser le module d'analyse. En effet, vu la longueur potentielle des dialogues (un dialogue peut durer plusieurs heures), il n'est pas impossible que le conducteur/chauffeur change en cours de dialogue, ou à cause d'une raison ou d'une autre, change sa langue de dialogue.

3.3.1.2 Le gestionnaire d'une Langue Particulière (GLP)

Ce module peut être vu comme une extension des gestionnaires de systèmes que nous avons utilisés dans les systèmes Corrector et Oasis. En effet, outre sa fonction de corridor d'information entre les différentes composantes du système, ce module est équipé d'un ensemble de règles qui lui permettent d'activer un sous-ensemble des unités d'analyse grammaticale et une seule unité d'analyse sémantique. Ces fonctions sont basées sur les attentes fournies par le GD aussi bien que les traitements de l'entrée effectués au sein de Navigator lui-même. Deux groupes de fonctions ont été implantés : un pour l'activation des unités d'analyse grammaticale et un pour l'activation d'une unité sémantique.

3.3.1.2.1 Les règles d'activation des unités syntaxiques

Comme la correspondance entre les unités de reconnaissances et les unités d'analyse syntaxique n'est pas directe, (parfois plusieurs unités de reconnaissance correspondent à une seule unité d'analyse syntaxique), nous avons utilisé des règles d'inférence spécifiques pour le routage des hypothèses de reconnaissance aux unités d'analyse grammaticales appropriées. Les règles utilisées prennent en considérations à la fois l'attente du GD associé à l'hypothèse et la tâche courante du dialogue. Voici un exemple d'une simplification d'une règle (en Prolog) utilisée pour la distribution des unités syntaxiques :

```
parse_unit_distribution(Input,city,route,Parse):-
parse_route(Input,Parse).
```

La règle précédente veut dire que si l'attente fournie par le GD correspond à un nom de ville et si la tâche courante est *route* alors l'entrée doit être envoyée à l'unité d'analyse grammaticale *route*.

3.3.1.2.2 Les règles d'activation des unités sémantiques

L'activation d'une unité sémantique se fait sur la base de trois critères : la nature des non-terminaux sémantiques de l'arbre Sm-TAG, la tâche courante et les attentes du GD. Ces trois critères sont combinés au sein de règles d'inférence dont la structure générale est similaire à celle utilisé pour le routage vers les unités grammaticales. En voici un exemple (en Prolog) :

```
frame_poi_route(Input,_Expectation,route,Frame):-
all_information_concepts(Input),
main_frame_information(Input,Frame).
```

La règle précédente, signifie que si la tâche courante est *route* et si tous les non-terminaux sémantiques de l'arbre d'analyse Sm-TAG sont tous propres au domaine *informations sur Vico* (cela se fait à l'aide d'une fonction spéciale qui scan l'arbre d'analyse) alors quelle que soit l'attente du GD cet arbre doit être envoyé à l'unité d'analyse sémantique du domaine *informations sur Vico*.

3.3.1.3 L'analyse grammaticale

Le module d'analyse de Navigator est basé sur le formalisme Sm-TAG. L'algorithme d'analyse utilisé est assez proche de celui utilisé dans Oasis. En effet, nous avons vu que la Sm-TAG est convertie en une combinaison de règles d'inférences et de RTRs. Cependant deux points distinguent l'analyse grammaticale de Navigator :

1. Implantation d'un compilateur Sm-TAG –RTRs pour faciliter l'écriture de la grammaire.
2. Division de la grammaire en différentes unités qui correspondent chacune à un domaine particulier de dialogue et qui partagent un ensemble de ressources grammaticales communes.

3.3.1.3.1 L'interface entre la grammaire et le module d'analyse

Pour automatiser la conversion des arbres Sm-TAG en RTRs nous avons implanté un module de compilation qui a pour entrée la grammaire au format Sm-TAG et dont la sortie est l'équivalent de cette grammaire au format interne du système d'analyse. Le schéma général de l'emplacement de ce module dans le système est le suivant :

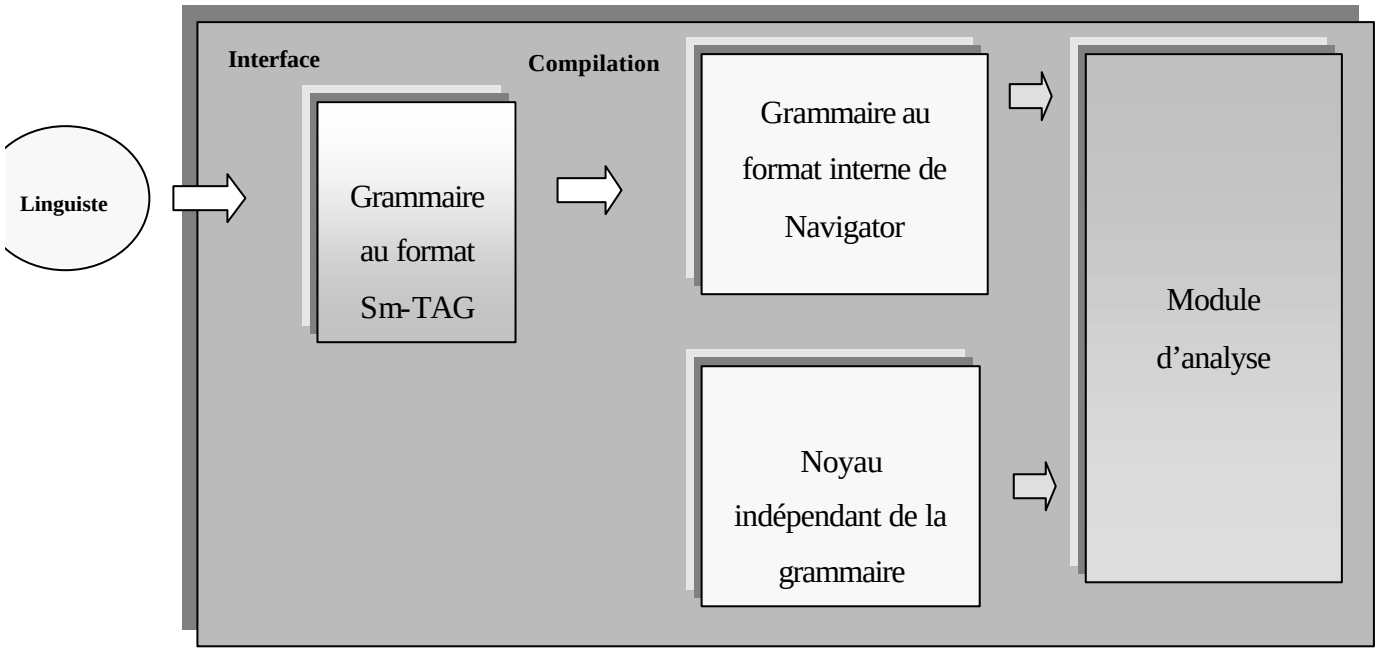


Figure 91. Le schéma général de l'emplacement du module de compilation de la grammaire au sein du système Oasis

Comme nous pouvons le voir dans le schéma précédent, la grammaire au format Sm-TAG est tout d'abord saisie par le linguiste et puis cette grammaire est compilée en un format interne qui, combiné au noyau indépendant de la grammaire, donne comme résultat le module d'analyse. Le noyau indépendant de la grammaire couvre un ensemble de principes généraux du formalisme qui ne dépendent pas d'une grammaire ou d'une application particulière (les règles d'inférence syntaxiques).

Ainsi, la compilation revient à convertir la grammaire Sm-TAG en un RTR étant donné que ce dernier intègre à la fois les arbres élémentaires et l'opération de substitution.

Avant de présenter les différentes étapes de traitement dans notre algorithme, nous allons commencer par une présentation des RTRs du point de vue implantation.

Un réseau de transition récursif est un graphe qui nécessite les informations suivantes :

1. Un dictionnaire qui contient tous les mots du lexique avec leurs catégories morpho-syntaxiques et/ou sémantiques.
2. Le nom du réseau ou sa catégorie principale.
3. L'état du commencement.
4. Une série d'états intermédiaires liés par des arcs étiquetés par des catégories dont la vérification constitue la condition nécessaire et suffisante pour le passage d'un état à l'état suivant.

5. L'état de la fin qui marque le succès du passage du réseau.

Ainsi, pour convertir les arbres du formalisme Sm-TAG en RTRs la procédure générale est la suivante :

1. Conversion des arbres lexicaux : convertir tous les arbres lexicaux t_l, en entrées lexicales dans le dictionnaire de la manière suivante :

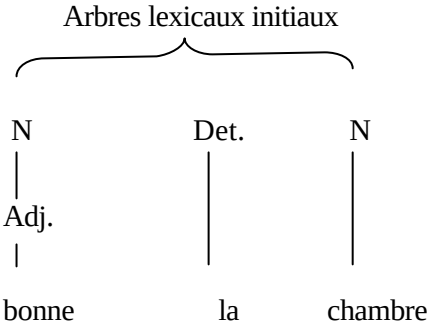
i- Soit *A* l'étiquette de la racine *m* de l'arbre t_l. Créer une nouvelle entrée lexicale et instancier la variable correspondant à la racine de l'arbre dans cette entrée par *A*.

ii- Si t_l est de profondeur deux et si l'étiquette du nœud interne de cet arbre est Y_i alors instancier la variable correspondant à cette information dans l'entrée lexicale par la catégorie Y_i. Sinon, si la profondeur de l'arbre est de un, alors remplacer cette variable par un élément vide.

iii- Soit Z_i l'ancre de l'arbre t_l (l'item lexical). Instancier le champ correspondant à l'item lexical dans l'entrée dans le dictionnaire par Z_i.

Ainsi, l'entrée dans le dictionnaire correspondant à un arbre lexical a la forme suivante : mot(racine_arbre, catégorie_morpho-syntaxique, ancre_lexical).

Pour rendre cette idée encore plus explicite, prenons les exemples suivants :



mot(n, adj, bonne).

mot(det, _, la).

mot(n, _, chambre).

Figure 92. Quelques arbres lexicaux et leur conversion en entrées lexicales du réseau de transition

Vue la simplicité du format des arbres lexicaux, ces arbres ont été saisis directement au format interne.

2. Conversion des arbres locaux et globaux : la conversion des arbres locaux et globaux *t* en réseaux de transition se fait de la manière suivante :

i- Soit *A* la racine *m* de l'arbre *t* définir *A* comme le nom du réseau *R*.

ii- Soit K_i, L_{i+1} , …, Z_n les nœuds fils de *A*.

iii- Créer les deux prédicats suivants : initial(i, A), final(n, A) (définition des états initial et final du réseau *R*).

iv- Créer les arcs du réseau de la manière suivante :

arc(i, i+1, k, A).

arc(i+1, i+2, l, A).

(….)

arc(n-1, n, z, A).

Pour concrétiser ces démarches, prenons à titre d'exemple les trois arbres élémentaires suivants et leurs équivalents en réseaux de transition :

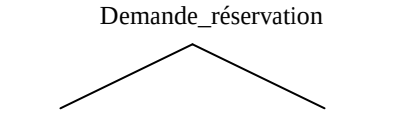
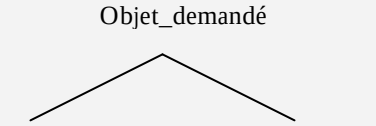
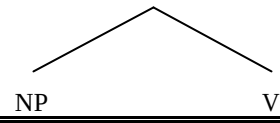
Les arbres locaux et globaux	**Les RTRs équivalents**
	initial(0, demande_réservation). final(2, demande_réservation). arc(0,1, formule_demande, demande_réservation). arc(1, 2, objet_demande, demande_réservation).
	initial(0, objet_demandé). final(2, objet_demandé). arc(0,1, chambre, Objet_demandé). arc(1, 2, Caractère_chambre, Objet_demandé).
	initial(0, formule_de_demande). final(2, formule_de_demande). arc(0,1, np, formule_de_demande). arc(1, 2, v, formule_de_demande).

Figure 93. Exemple d'arbres élémentaires locaux et globaux et les RTRs équivalents

Comme nous pouvons le constater dans la figure précédente, les réseaux produits correspondent parfaitement aux arbres locaux et globaux donnés.

Pour l'écriture des arbres locaux et globaux nous avons défini un format spécial dont le schéma général est le suivant :

(Racine_arbre,

[nœud_fils$_1$,

nœud_fils$_2$, …,

nœud_fils$_n$],

Commentaires).

Par ailleurs, le compilateur est équipé d'une fonction qui permet de compiler différentes grammaires (pour les dialogues multi-domaines) qui doivent être mises dans des fichiers séparés.

3.3.1.3.2 La modularité de la grammaire

La spécificité principale de Navigator par rapport à Oasis est sa modularité. En effet, nous avons organisé la grammaire de manière à refléter l'organisation des domaines du dialogue. Voici les principaux traits liés à la modularité de la grammaire :

I. Les unités utilisées : la grammaire est divisée en un ensemble de parties qui correspondent chacune à un domaine de dialogue et qui partagent un ensemble d'arbres indépendants du domaine. Ainsi, nous avons obtenu cinq unités d'analyse grammaticale plus l'unité des ressources communes :

1. **Grammaire pour la tâche d'informations routières :** cette grammaire couvre principalement des énoncés de demande de déplacement vers une adresse particulière : région, ville, rue, etc.
2. **Grammaire de point d'intérêt :** cette grammaire est complémentaire de la précédente dans la mesure où elle est destinée à traiter des énoncés de demande vers des points d'intérêts. La liste des points d'intérêt qui ont été retenus par le consortium englobe vingt-cinq types dont : hôtels, hôpitaux, gares, boîtes de nuits, etc.
3. **Réservation hôtelière :** cette grammaire est destinée à couvrir différents types d'énoncés liés à la réservatoin d'une chambre d'hôtel comme les demandes de réservation, les caractéristiques principales des chambres et des hôtels, les dates d'arrivée et de départ, etc.
4. **Informations sur Vico (aide) :** cette grammaire est destinée à couvrir des énoncés de demande d'information sur le système Vico, sa technologie, ses fonctionnalités, etc. Nous avons implanté une grammaire pour l'anglais seulement puisque cette tâche ne fait pas partie officiellement des tâches du projet. L'implantation de cette grammaire pour l'anglais avait pour but de tester l'utilité de l'ajout d'une telle tâche au système.
5. **Module de traitement des épellations :** ce module n'est pas équipé d'une grammaire spécifique mais il est équipé d'une interface qui assemble les lettres et les envoie à l'unité d'analyse grammaticale choisie par le GLP.
6. **Les ressources communes :** il s'agit d'un ensemble d'arbres qui ne dépendent pas d'une application particulière et qui par conséquent peuvent être utilisés par les différentes unités. Les formules de demande *I want to*, les formules de confirmation *yes* et les demandes de répétition *repeat again* sont des exemples de segments correspondants à des arbres partagés entre les différents domaines.

II. Motivations du partage en unités séparées : différentes raisons ont motivé notre choix d'une approche modulaire pour l'analyse grammaticale :

Simplification des procédures d'écriture et maintien des grammaires : la division de la grammaire en différentes parties qui correspondent chacune à un domaine applicatif différent rend possible le partage de l'effort d'écriture des grammaires à plusieurs linguistes qui peuvent travailler en parallèle sur la grammaire. Par ailleurs, cela facilite le maintien de la grammaire dans la mesure où il est plus facile d'intégrer de nouvelles règles dans une grammaire dont la taille est relativement petite que dans une grande grammaire.

1. **Augmentation de la rapidité de traitement :** la rapidité de traitement est une propriété désirée de tous les systèmes de compréhension du langage oral étant donné que ces systèmes sont destinés à fonctionner en ligne. La rapidité d'un algorithme est mesurée par deux formules : $O(n^x)$ où n est la longueur de l'entrée et $O(G^x)$ où G est la taille de la grammaire. Comme le note (Joshi,1996), la complexité réelle d'un algorithme est souvent inférieure à sa complexité théorique et ce selon la taille de G. Etant donné que la division de la grammaire affecte seulement les arbres qui se combinent avec l'opération de substitution (comme nous avons dit, les arbres lexicaux initiaux et secondaires ne sont pas partagés puisqu'ils ne sont dépendants du domaine), la valeur de X est 2. Ainsi, si nous divisons une grammaire en quatre parties de tailles égales la complexité spaciale sera quatre fois moins grande : $O((G/4)^2)$.
2. **Réduction des ambiguïtés :** en réduisant l'espace de recherche à l'espace jugé pertinent par rapport à l'énoncé d'entrée, nous réduisons aussi les ambiguïtés liées au traitement. Par exemple, dans le contexte d'information routière, les numéros utilisés correspondent seulement à des numéros de rues comme dans : *five Cikorievej please* ou *it is five, the number is five,* etc. Par contre, si nous considérons la totalité de la grammaire, les numéros peuvent référer à différents objets comme le nombre de personnes, le prix, le nombre des chambres demandées, etc. Ainsi, la modularité de la grammaire permet dans certains cas d'**éviter** les ambiguïtés plutôt que d'avoir à les résoudre.

3.3.1.4 Le module d'arbitrage

Le module d'arbitrage est un classifieur multicritères qui a pour fonction de sélectionner la meilleure analyse parmi les N analyses reçues du module d'analyse grammaticale. Les critères retenus pour la classification couvrent pratiquement tous les niveaux des connaissances manipulés par Vico : connaissances acoustiques, connaissances syntaxiques et sémantiques, connaissances pragmatiques et connaissances sur le domaine. Ainsi, si nous utilisons la terminologie de la théorie de l'optimalité (voir (Gilbers et De Hoop, 1998) pour une introduction à cette théorie), l'énoncé à retenir est celui qui satisfait au maximum les contraintes précitées. Le problème, est que vue la complexité et l'hétérogénéité de ces contraintes, il existe différents conflits de priorité entre ces contraintes qui sont à résoudre. Ainsi, dans notre module d'arbitrage nous combinons des

principes génériques dérivés des différentes sources de connaissances précitées à des paramètres empiriques qui reflètent le poids de chacune de ces sources dans le traitement. D'un point de vu formel, le score global devrait se calculer selon la formule suivante :

$$\sum S_{Ti}(H) \,.\, P_{Ti}$$

Où :

1. S_T est une fonction de score partiel de *T* qui est un type de connaissance quelconque (connaissances acoustiques, connaissances pragmatiques, etc.).
2. *H* est une hypothèse de reconnaissance.
3. *n* est le nombre des types de connaissance.
4. P_T est le poids du type *T* dans le traitement.

En pratique, la situation est plus complexe vu les interdépendances entre les différentes sources de connaissances qui doivent être prises en condiération lors du calcul. Ainsi, nous allons procéder en deux étapes :

i. Création de deux scores initiaux : un pour les connaissances perceptives que nous avons appelé score global de reconnaissance (le score de reconnaissance pondéré sémantiquement et pragmatiquement) et un pour l'analyse grammaticale.

ii. Combinaison de ces scores en un score global de l'énoncé qui sera utilisé pour sélectionner la meilleure analyse.

Dans notre démarche de calcul nous avons utilisé un bon nombre de poids de natures diverses. Notons que les valeurs de ces poids ont été trouvées empiriquement à l'aide d'un corpus de quarante hypothèses de reconnaissance.

3.3.1.4.1 Le score global de reconnaissance

Comme nous avons vu, les systèmes de reconnaissance produisent *N* graphes de mots contenant chacun une liste de mots associés à leurs scores de reconnaissance. Pour calculer un score global de reconnaissance de chaque hypothèse à partir des scores individuels, nous procédons selon les étapes suivantes :

1. **Pondération sémantique :** il s'agit de distinguer entre les degrés d'importance de l'information transmise par les mots des points de vues linguistiques et pragmatiques en leur associant un score qui reflète cette importance. Ainsi, nous avons distingué entre trois classes de mots :
   i. Les mots qui ont un contenu directement important : cette liste englobe tous les mots qui permettent de remplir un slot dans le schéma sémantique comme les noms géographiques (nom de ville, de rue, etc.), les nombres (nombre de personnes, nombre de chambres, etc.).

ii. Les mots qui jouent un rôle important linguistiquement : il s'agit des mots qui peuvent être la tête de syntagme comme les verbes, noms, adverbes, etc.

iii. Le reste.

A chacune de ces catégories, nous avons associé un poids P_x (où $0< P \leq 1$ et x est l'une des trois catégories précitées).

2. **Combinaison des scores individuels :** la combinaison des scores individuels pondérés sémantiquement se fait selon la formule suivante :

$$\text{Score combiné} = \frac{\sum S_X(H)}{\sum P_X(H)}$$

Où Sx est le score de reconnaissance d'un mot *X* de l'hypothèse *H et* P_X est le poids associé au score de *X*.

Cette formule permet, en cas de différence significative des scores de reconnaissance des mots, de produire un score combiné qui reflète les scores des mots les plus importants sémantiquement dans l'hypothèse. Par ailleurs, en cas d'égalité des scores de reconnaissance le score combiné obtenu est la moyenne des scores individuels avant la pondération sémantique (dans ce cas les poids sémantiques ne sont pas très utiles).

3. **Pondération pragmatique du score combiné :** il s'agit d'associer un poids à chaque hypothèse qui correspond au degré d'attente de cette hypothèse par le gestionnaire de dialogue. Sur ce plan, nous avons distingué entre deux types d'énoncé associés à deux poids différents :

i. Les énoncés informatifs : il s'agit d'énoncés dont le contenu sémantique est directement lié à l'exécution de la tâche (questions, réponses, assertions, etc.). Un poids normal à ces énoncés.

ii. Les énoncés qui correspondent à des méta-communications (demande de répétition par exemple). Le poids associé aux énoncés de ce type est moins important que celui associé aux énoncés précédents.

Le résultat de la pondération pragmatique est le Score Global de Reconnaissance (SGR).

3.3.1.5 Le score d'analyse grammaticale

Etant donné que nous avons adopté une approche d'analyse combinant une analyse superficielle à une stratégie sélective, il est nécessaire d'avoir des critères qui permettent de juger la qualité de l'analyse produite par le module grammaticale. Ainsi, le score d'analyse grammaticale peut-être vu comme un moyen pour pénaliser les arbres dont l'obtention a nécessité le recours à l'une de nos approches de relaxation : l'analyse partielle ou la stratégie sélective. Le score de l'analyse grammaticale (qui a généralement la forme d'un ensemble de segments ou d'îlots) est calculé de la manière suivante :

1. Classification des segments : nous avons distingué entre deux types de segments : les segments non-analysés et les segments analysés. Chacun des segments d'une analyse est associé à un score qui reflète sa taille. Ce score est appelé : Score de Couvertur de Segment (SCS). Le SCS d'un segment non-analysé est le nombre des mots (non-analysés) que couvre ce segment alors que celui d'un segment analysé correspond au nombre des arbres locaux et globaux dominés par la racine de ce segment.
2. Calcul des Scores Locaux de l'Analyse Grammaticale (SLAG) : le SLAG d'un segment est calculé de la manière suivante : $SLAG_X = CCS_X . P_X$ où *X* est un segment quelconque et P_X est le poids de *X* dans le traitement. Il faut noter que le poids varie selon la nature du segment. Ainsi, nous associons un poids positif aux segments analysés alors que nous associons un poids négatif aux segments non-analysés. Ainsi, plus le nombre des arbres locaux et globaux compris dans un segment analysé est élevé plus le score de ce segment est élevé. Par contre, plus le nombre des mots non-analysés est élevé plus le score de ce segment est bas.
3. Calcul du Score Global de l'Analyse Grammaticale (SGAG) : le SGA reflète le score global de l'analyse associée à l'énoncé. Afin de refléter la qualité de l'analyse d'une entrée donnée, ce score doit prendre en considération non seulement les scores locaux de ses segments mais aussi la longueur de l'hypothèse. Cela permet de favoriser les analyses dont les segments couvrent plus de mots. Voici la formule que nous avons adoptée pour le calcul du SGAG :

$$SGA = \frac{\sum SLAG_X}{P . L}$$

Où *P* est un poids dont la valeur est trouvée empiriquement et *L* est la longueur de l'hypothèse.

3.3.1.5.1 Calcul du Score Global de l'Enoncé (SGE)

Le SGE est obtenu en calculant la moyenne du SGR et du SGAG. Il faut noter que la valeur de *P* (le poids empirique) dans la formule utilisée pour le calcul du SGAG peut être considérée comme un poids qui détermine son importance dans le SGE.

3.3.1.5.2 Calcul du score normalisé

Le GD dispose de trois stratégies dont le choix dépend du SGE fournit par le module d'arbitrage. Afin de rendre ce score utilisable par le GD, nous avons implanté une fonction de normalisation qui concrétise ses valeurs. Ainsi, les scores SGE dont la valeur est inférieure à 0.25 sont remplacés par la valeur 1 (mauvais). Les SGE dont la valeur se situe entre 0.25 et 0.5 sont remplacés par la valeur 2 (moyen). Finalement les scores dont la valeur est située entre 0.5 et 1 sont remplacés par la valeur 3 (bon).

3.3.1.6 L'analyse sémantique

Le module d'analyse sémantique a pour fonction de convertir les arbres d'analyse grammaticale (qui combinent la syntaxe et la sémantique) en une représentation sémantique pure utilisable directement par le gestionnaire de dialogue.

D'un point de vue formel, un schéma est défini par un nom et un ensemble d'attributs (*Slots*). Chaque schéma est implanté comme une unité indépendante dont l'activation se fait selon différents critères (comme nous avons vu avec le GLP). Dans Navigator, nous avons utilisé quatre unités d'analyse sémantique qui partagent un ensemble de ressources communes. Il s'agit de l'unité de navigation routière, l'unité de points d'intérêts, l'unité de réservation hôtelière et l'unité d'information (aide), les ressources partagées : il s'agit d'un ensemble de slots communs à tous les schémas comme la confirmation, la négation, la demande de répétition, etc.

L'analyse sémantique se fait par un ensemble de règles d'inférence dont la fonction est d'extraire les informations pertinentes pour le schéma à partir des arbres d'analyse.

En voici un exemple simplifié :

analyse_pois_location([gtr_location_street1,_,[No,_],[Name,_]],[[street_name,Name],[street_number,NO]]).

La règle précédente permet d'extraire le numéro et le nom de la rue de l'arbre d'analyse Sm-TAG et d'utiliser les valeurs de ces deux variables pour instancier les slots appropriés.

Le module d'analyse sémantique que nous avons implanté produit uniquement les slots dont les valeurs sont instanciées. Un module de post-traitement (le module d'enveloppe) que nous allons présenter plus loin effectue un formatage de la sortie du module d'analyse sémantique et l'enrichit avec les slots non-instanciés.

3.3.1.7 Le module de traitement des extragrammaticalités

Nous avons intégré notre module Corrector pour le traitement des extragrammaticalités au sein de Navigator. Le développement de modules similaires à Corrector pour l'allemand et l'italien est prévu pour la deuxième phase du projet.

3.3.2 Exemple de traitement

Pour donner une idée sur les différentes étapes de traitement dans le cadre du système Navigator prenons l'exemple suivant[51] : *I want to go to Trento*.

1. **Entrée :** l'énoncé précédent est traité par le module de reconnaissance qui produit une liste de deux hypothèses qui correspondent à deux unités de reconnaissance actives. La sortie du système de reconnaissance est enrichie par le GS qui y ajoute l'information sur la langue courante ainsi que la tâche courante de dialogue (reçue du GD). Ainsi, l'entrée de Navigator (reçue par le GGN) a la forme

[51] Dans cet exemple, nous avons procédé à certaines simplifications et changements de format afin de clarifier le propos.

suivante : current_language(english) current_task(route) SRU_0 I 0.55 want 0.64 to 0.39 go 0.5 to 0.6 Trento 0.66 SRU_3 yes 0.21 #noise# 0.18 #noise# 0.23

2. Le GGN active le module de l'anglais et envoie la chaîne *current_task(route) SRU_0 I 0.55 want 0.64 to 0.39 go 0.5 to 0.6 Trento 0.66 SRU_3 yes 0.21 #noise# 0.18 #noise# 0.23* à son GLP.

3. **Analyse grammaticale :** le GLP de l'anglais extrait les mots des deux hypothèses. La première hypothèse (I want to go to Trento) est envoyée à l'untié d'analyse grammaticale numéro 1 (spécialisé dans la navigation routière). Cette unité est choisie parce que la tâche courante de dialogue est *route* et en même temps l'attente du GD associée à cette hypothèse est *route* (l'attente du GD est déduite du tag de l'unité de reconnaissance : SRU_0). Puis, la même chose est répétée avec la deuxième hypothèse. Ainsi, la sortie du module d'analyse grammaticale est une pile qui contient les analyses correspondantes aux deux hypothèses :

[

[[route_global, [request_formulation1, [pron_subj, I], [verb, want], [preposition, to]], [requested_object_route, [verb, go]], [destination_city, [preposition, to], [proper_name, Trento]]]]**,**

[[non-analyse, [#noise#, noise]], [non-analyse, [#noise#, noise]], [confirmation_simple, [adverb, yes]]]]

]

4. **Arbitrage :** la sélection de la meilleure des deux analyses reçues par le module d'arbitrage se fait de la manière suivante :

iv. Calcul du score global de reconnaissance : le score global de reconnaissance est calculé de la manière suivante :

a. Pondération sémantique : les scores individuels des mots de chacune des deux hypothèses sont pondérés par un poids qui reflète l'importance de l'information qu'ils transmettent : SRU_0 I (0.55 × 0.25 = 0.13) want (0.64 × 0.5 = 0.32) to (0.39 × 0.25 = 0.09) go (0.5 × 0.5 = 0.25) to (0.6 × 0.5 = 0.3) Trento (0.66 × 0.75 = 0.495) SRU_3 yes (0.21 × 0.75 = 0.15) #noise# (0.18 × 0.25 = 0.04) #noise# (0.23 × 0.25 = 0.05).

b. Calcul du Score de Reconnaissance Combiné (SRC) des deux hypothèses : cela se fait de la manière suivante : SRC_0 = 0.13 + 0.32 + 0.09 + 0.25 + 0.3 + 0.495 (somme des scores pondérés = 1.58) / 0.25 + 0.5 + 0.25 + 0.5 + 0.5 + 0.75 (Somme des poids sémantiques = 2.75) = 0.57 ; SRC_3 = 0.15 + 0.04 + 0.05 / 0.75 + 0.25 + 0.25 = 0.19.

c. Calcul du Score Global de Reconnaissance (SGR) : le SGR est obtenu en pondérant le SRC de chaque hypothèse sémantiquement. Ainsi, les SGRs des deux hypothèses sont obtenus de la manière suivante : SGR_0 = 0.57 × 1 = 0.57 ; SGR_3 = 0.19 × 0.7 = 0.11.

ii. Calcul du SGAG : le Score Global d'analyse Grammatical est obtenu selon les étapes suivantes :

a. Classification des segments : dans cette étape, les deux arbres d'analyse grammaticale sont convertis en un format standard. Ainsi, nous obtenons : [SRU_0, (parsed, 3)] et [SRU_3, (unparsed, 2), (parsed, 1)].

b. Calcul des Scores Locaux de l'Analyse Grammaticale (SLAG) : [SRU_0, (parsed, $3 \times 1 = 3$)], [SRU_3, (unparsed, $2 \times -1 = -2$), (parsed, $1 \times 1 = 1$)].

c. Calcul du Score Global de l'Analyse Grammaticale (SGAG) : $SGAG_0 = 3 / (1.5$ (poids empirique) $\times 6) = 0.3$; $SGAG_0 = -1 / (1.5 \times 3) = -0.22$.

iii. Calcul du SGE et sélection de la meilleure hypothèse : les scoer glaobaux des deux hypothèses sont calculés de la manière suivante : $SGE_0 = 0.57 + 0.3 / 2 = 0.43$; $SGE_3 = 0.11 + (-0.22) / 2 = -0.05$. L'hypothèse fournie par l'unité SRU_0 est celle qui est retenue comme son score global est supérieur à celui de l'hypothèse fournie par SRU_3.

iv. Normalisation du SGE : le score de l'hypothèse retenue est normalisé. Comme sa valeur se situe entre 0.25 et 0.5 il est remplacé par le score 2.

5. **Activation de l'unité sémantique appropriée :** comme la tâche courante est route et l'attente du GD associée à l'hypothèse retenue par le module d'arbitrage est aussi route, l'analyse grammaticale de l'hypothèse retenue est envoyée au module d'analyse sémantique spécialisé dans la tâche d'informations routières.

6. **Analyse sémantique :** tout d'abord, le système ouvre une entête d'un schéma pour l'information routière auquel il associe le score normalisé comme un slot. Puis l'arbre d'analyse grammaticale est parcouru de droite à gauche par l'algorithme d'analyse sémantique. Le module localise un sous-arbre marqué pour la propagation (destination_city), une règle d'inférence est utilisée pour extraire le slot city_name de cet arbre. La sortie finale du module d'analyse sémantique est la suivante : [route_frame,[city_name,Trento],[confidence_score, 2]].

3.3.3 Discussion de l'architecture de Navigator

3.3.3.1 Aspects logiciels

Le choix de l'architecture en général est motivé par les mêmes raisons que celles d'Oasis et Corrector. Cependant deux spécificités de l'architecture de Navigator méritent d'être citées :

1. **Modularité :** comme nous avons vu, Navigator a été construit sur le principe de maximiser le partage des ressources logicielles et linguistiques entre les langues et entre les applications. Cela facilite considérablement le développement et le maintien du système puisqu'on maximise l'utilisation des ressources déjà existantes. Par ailleurs, nous avons vu que la modularisation de la grammaire a des

avantages en terme de rapidité de traitement qui est un trait important pour un système de compréhension.

2. **Souplesse :** l'utilisation d'une interface unique pour les modules des trois langues a facilité l'intégration de Navigator avec le reste des modules de Vico. En effet, l'effort d'intégration à été réduit à l'intégration d'un seul module plutôt que trois.

3.3.3.2 Aspects cognitifs

D'un point de vue cognitif, deux points clés de l'architecture peuvent être notés :

1. **Guidage du module d'analyse grammaticale par les attentes du GD :** sur le plan cognitif, l'architecture de Navigator peut être vue comme une extension du principe de l'interaction entre les différents niveaux de connaissance sur lequel est fondé la Sm-TAG. Ainsi, grâce à la prise en considération des attentes dialogiques (qui ne peuvent pas être prise en considération directement par la Sm-TAG), l'architecture de Navigator permet de guider le module d'analyse grammaticale selon les attentes de haut niveau. Comme nous l'avons vu précédemment, ce mode de guidage a été relevé dans différents travaux de psycholinguistique expérimentale comme ceux de (Spivey-Knowlton, 1994) et (Boland *et al.*, 1995).

2. **Perception et compréhension au sein de Navigator :** il est couramment admis dans la communauté de psycholinguistique expérimentale (voir (Schwartz, 1996), (Kurdi, 1996) pour une revue de différents travaux dans ce domaine) que la perception n'est pas entièrement indépendante de la compréhension. En effet, l'identification des phonèmes et la combinaison de ces phonèmes en mots dépend non seulement des facteurs acoustiques mais aussi de différents facteurs tant linguistiques (comme la phonologie, la morphologie, la syntaxe et la sémantique) que pragmatiques. La décision de l'hypothèse à retenir (décision perceptive) est faite après l'analyse grammaticale de toutes les hypothèses perceptives et le choix de la meilleure hypothèse est fait selon pratiquement toutes les sources d'informations impliquées dans la perception et la compréhension de la parole.

3.3.4 Réalisation du système Navigator

D'un point de vue pratique, nous pouvons distinguer entre deux étapes dans la réalisation du système Navigator : l'écriture des grammaires et l'implantation du système.

3.3.4.1 Les grammaires utilisées

3.3.4.1.1 Le corpus utilisé pour l'écriture de la grammaire

Les grammaires écrites ont été basées sur trois corpus qui ont été collectés dans trois sites : (Nislab-Odense pour l'anglais, Bosch-Stuttgart pour l'allemand, IRST-Trento pour l'italien). Un protocole commun a été adopté pour la collecte dans les trois sites. Les propriétés principales de la collecte des données sont décrites dans les points suivants :

1. Adoption de la méthode de magicien d'Oz : les sujets devaient dialoguer avec un humain via une interface logicielle qui leur est présentée comme le système de dialogue. Cela permet d'avoir un comportement dialogique proche de celui que les sujets auraient adopté en cas de dialogue avec un système de dialogue réel.
2. Simulation d'une situation de conduite : les sujets devaient répondre aux questions du magicien tout en conduisant un simulateur de voiture. Les simulateurs utilisés consistent en un jeu vidéo de conduite de voiture dont la commande se fait de manière proche de celle des voitures réelles (avec un volant et deux pédales). L'objectif de cette utilsation est de simuler la charge cognitive de la conduite et son effet potentiel sur le dérouelement de dialogue.
3. Sept scénarios qui couvrent les différents domaines de dialogue ont été utilisés pour la collecte des données. Les scénarios portent sur des demandes dans la région du Trentino en Italie.
4. Des locuteurs natifs ont été utilisés pour la collecte de l'italien et l'allemand alors que les locuteurs utilisés pour la collecte de l'anglais étaient des danois qui maîtrisent l'anglais.

Le résultat de cette collecte de données en termes d'énoncés d'utilisateur (qui sont utilisés pour l'écriture de la grammaire) est le suivant : 1220 énoncés pour l'allemand (1004 ont été utilisés pour l'écriture et l'évaluation de la grammaire), 1067 (886 énoncés ont été utilisés pour l'écriture de la grammaire) et 942 énoncés pour l'italien (dont 855 ont été utilisés pour l'écriture de la grammaire.

Outre les données obtenues avec la simulation de magicien d'Oz, un corpus de 180 énoncés qui portent sur des expressions temporelles (dates, expression d'arrivée ou de départ) a été obtenu pour l'anglais en simulant des énoncés. Le corpus a été obtenu en utilisant une interface développée avec Power Point. Les principales informations affichées dans les transparents sont les suivantes :

1. Le résumé de l'historique d'un dialogue virtuel (par exemple, vous avez demandé de réserver une chambre simple).
2. L'énoncé du système qui demande l'expression temporelle (vous arrivez quand ? Vous partez quel jour?, etc.).
3. Des instructions sur le contenu de la réponse à donner (indiquer le jour et le mois de votre arrivée, indiquer le mois et l'année de votre départ, etc.).

Ce corpus a été ensuite traduit et augmenté par des locuteurs natifs en allemand et en italien afin d'enrichir les grammaires des expressions temporelles dans ces deux langues.

3.3.4.1.2 Ecriture de la grammaire

Notre contribution à l'écriture des grammaires de Navigator est présentée dans les points suivants :

1. **La grammaire de l'anglais :** la tâche d'écriture de la grammaire pour l'anglais a été effectuée par nous.

2. **La grammaire de l'italien :** nous avons formé la linguiste qui a écrit la partie principale de la grammaire, supervisé son travail et écrit environ 20% des arbres locaux et globaux.
3. **La grammaire de l'allemand :** en ce qui concerne l'allemand notre rôle s'est limité à former et superviser les deux linguistes qui ont travaillé successivement sur cette grammaire.

Par ailleurs, nous avons écrit toutes les règles d'inférences sémantiques pour les grammaires des trois langues. Les grammaires obtenues sont décrites dans le tableau suivant :

Langue	No. des arbres locaux et globaux	No. des arbres lexicaux généraux	No. des arbres lexicaux des noms de lieux
Anglais	456	989	16315
Allemand	671	1342	
Italien	308	953	

Tableau 26. Les tailles des grammaires écrites dans le cadre du système Navigator

Comme nous pouvons le remarquer dans le tableau précédent, les trois grammaires ont des tailles différentes que ça soit en terme d'arbres locaux et globaux d'une part qu'en terme d'arbres lexicaux d'autre part. Cette différence est due principalement aux différences linguistiques entre les trois langues du projet ainsi qu'à la différence des trois corpus utilisés pour l'écriture des trois grammaires.

3.3.4.2 Description des modules implantés

Tout comme avec nos systèmes précédents (Corrector, Safir et Oasis), nous avons utilisé le langage Prolog pour implanter le système Navigator[52].

Les modules et programmes réalisées dans le cadre du système Navigator peuvent être divisés en deux parties : les modules dépendants de la langue et les modules indépendants de la langue.

3.3.4.2.1 Implantation des modules dépendants de la langue

Il s'agit des trois modules d'analyse grammaticale, des trois GLPs, des trois modules d'analyse sémantique et des trois modules de calcul de score d'analyse syntaxique. En ce qui concerne l'analyse grammaticale nous avons distingué entre la grammaire Sm-TAG codée en Prolog selon le format que nous avons présenté plus haut et la grammaire compilée sem-automatiquement.

[52] La totalité des modules de Navigator que nous aovons présenté ont été développé par nous. Seul le module d'eveloppe que nous allons présenter plus loin a été développé par l'un de nos collègues du NISLab.

Les détails de l'implantation des modules spécifiques à l'anglais sont présentés dans le tableau suivant :

Module	Sous-modules	No. fichiers	No. lignes
GLP anglais	-	1	432
Analyseur	Grammaire Sm-TAG	4	6355
	RTRs Générés automatiquement par le compilateur à partir de la grammaire Sm-TAG	3	8450
Schéma	Unités de schémas	4	2169
	Ressources communes	1	342
Arbitrage	Calcul du score d'analyse syntaxique	1	648

Tableau 27. Les détails sur l'implantation de la version anglaise dus système Navigator

Sont présentés dans le tableau suivant les détails de l'implantation des modules spécifiques à la langue allemande :

Module	Sous-modules	No. fichiers	No. lignes
GLP allemand	-	1	327
Analyseur	Grammaire Sm-TAG	3	6838
	RTRs Générés automatiquement par le compilateur à partir de la grammaire Sm-TAG	3	11599
Schéma	Unités de schémas	3	2128
	Ressources communes	1	295
Arbitrage	Calcul du score d'analyse syntaxique	1	850

Tableau 28. Les détails sur l'implantation de la version allemande du système Navigator

Une description de l'implantation des modules spécifiques à la langue italienne est faite dans le tableau suivant :

Module	Sous-modules	No. fichiers	No. lignes
GLP italian	-	1	293
Analyseur	Grammaire Sm-TAG	3	4133
	RTRs Générés automatiquement par le compilateur à partir de la grammaire Sm-TAG	3	5712
Schéma	Unités de schémas	3	1873
	Ressources communes	1	284
Arbitrage	Calcul du score d'analyse syntaxique	1	452

Tableau 29. Les détails sur l'implantation de la version italienne dus système Navigator

3.3.4.2.2 Implantation des modules indépendants de la langue

A leur tour les modules indépendants de la langue peuvent être divisés en deux parties : les modules utilisés dans le traitement et les modules utilisés directement dans le traitement et les modules fonctionnels.

1. **Les modules utilisés dans le traitement :** il s'agit de deux composantes du module d'arbitrage, d'un programme spécifique pour le traitement des graphes de mots reçus des systèmes de reconnaissance ainsi que le lexique géographique. Les détails de ces modules sont présentés dans le tableau suivant :

Module	Sous-modules	No. fichiers	No. lignes
Arbitrage	Traitement des scores de reconnaissance	1	311
	Combiaison des scores et sélection de la meilleure hypothèse	1	73
-	Segmentatoin des graphes de mots et extraction des scores de reconnaissance	1	204
	Lexique géographique (généré automatiquement à partir de la base de données)	1	16445

Tableau 30. Les ressources communes entre les trois langues

2. **Les modules fonctionnels :** il s'agit des modules utilisés hors-ligne pour la compilation de la Sm-TAG et la génération des arbres et patrons lexicaux correspondants au lexique géographique. Une description

générale de ces deux modules ainsi que la taille de l'implantation en terme de lignes de code sont présentés dans le tableau suivant :

Module	Fonction	No. lignes
Compilateur	Compile les arbres Sm-TAG en RTRs, génère des prédicats et produit un exécutable qui permet de tester directement la grammaire compilée.	780
Génération lexicale	Génère des entrées lexicales correspondants aux items de la base de données géographique. Les entrées lexicales générées sont ou bien des arbres lexicaux simples ou des patrons lexicaux correspondants aux noms de lieu qui comptent plus d'un mot.	328

Tableau 31. Les outils secondaires développés dans le cadre du projet Vico

3.3.4.3 Le module d'enveloppe

Comme nous avons vu dans les paragraphes précédents, le système Navigator a été implanté en Prolog que nous avons choisi pour différentes raisons dont les principales sont son adaptation au traitement automatique du langage Naturel et la rapidité du développement possible avec ce langage. Pour permettre à Navigator de communiquer avec le reste des modules de Vico un modules d'enveloppe a été développé.

L'interaction entre d'une part Navigator et le module d'enveloppe et d'autre part le module d'enveloppe et le GS de Vico sont présentés dans la figure suivante :

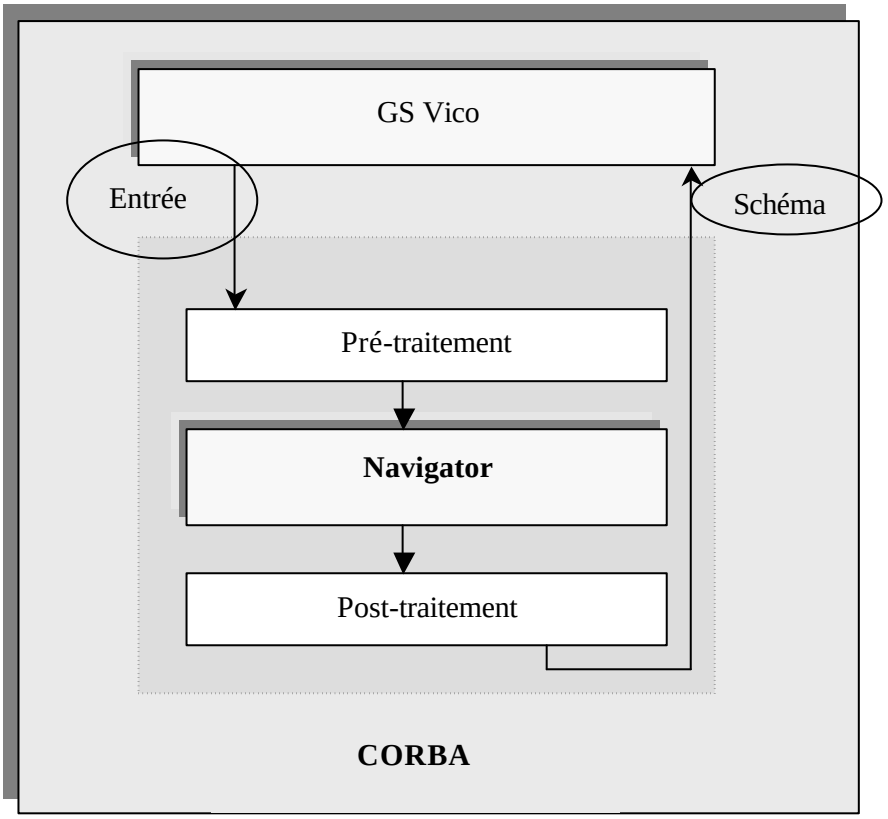


Figure 94. Schéma général du module d'enveloppe

Comme nous pouvons le voir dans la figure précédente, toutes les communications de Navigator avec le reste des modules se font à travers le module d'enveloppe.

Deux raisons principales ont motivé notre implantation de ce module :

1. **Interaction avec l'environnement CORBA :** comme il n'existe pas un compilateur opérationnel de l'IDL (*Interface Description Language*) de CORBA (qui est le protocole adopté par le consortium du projet Vico pour l'interaction des différents moduoles) pour la version de Prolog que nous avons adopté pour notre implantation (Swi-Prolog), nous utilisons le module d'enveloppe écrit en C++ comme un intermédiaire logiciel entre l'environnement CORBA et Navigator.
2. **Formatage des schémas produits par Navigator :** les schémas produits par Navigator sont des listes de mots selon la syntaxe de Prolog. Cette liste est considérée comme une chaîne de caractères selon le langage C++ (utilisé pour l'implantation du GD). Ainsi, le module d'enveloppe (le post-traitement) converti le schéma du format Prolog au format C++ (Objet schéma). Outre les aspects purement logiciels (comme le changement de la définition des entiers du type *caractère* en *entier*), le module d'enveloppe enrichit le schéma obtenu avec les slots non-instanciés qui sont utilisés par le GD comme base de raisonnement. Par exemple, lorsque ce module reçoit le schéma suivant fournit par Navigator : [route_frame,[city_name,Trento],[confidence_score, 2]], il fournit un schéma final qui a la forme suivante :

Type	route
PartOfCountry	
City	Trento
PartOfCity	
Street	
Number	
ConfScore	2
TypeMismatch	0
LexAmbiguity	0
POI Item[0]	
POI Item[1]	
POI Item[2]	
POI Item[3]	
POI Item[4]	
Unique	-1
Complete	-1
NonUniqueMax3	0
NonUniqueMore3	0
Inconsistent	-1
StreetNumbersDB	-1

Figure 95. Exemple de schéma fourni par le module d'enveloppe

Finalement, il n'est probablement pas inutile de mentionner que Navigator a été intégré avec succès au reste du système Vico et ce pour la première démonstration du projet qui a eu lieu le 28 novembre 2002.

3.3.5 Première evaluation de l'analyse linguistique dans Navigator

3.3.5.1 Objectif de l'évaluation

Après avoir évalué l'adaptation de la Sm-TAG au traitement des principaux phénomènes grammaticaux avec le système Oasis, nous avons l'intention dans cette évaluation de tester les traits propres à Navigator. Ainsi,

nous allons nous concentrer principalement sur l'adaptation de la Sm-TAG au traitement dans le contexte de systèmes de dialogue multi-domaine.

3.3.5.2 Matériel utilisé pour l'évaluation

Afin de concentrer notre évaluation sur l'adaptation de la Sm-TAG au traitement des dialogues multi-domaines, nous avons jugé bon de neutraliser les différentes variables non-pertinentes pour notre objectif en utilisant un matériel de test qui a les propriétés suivantes :

1. Nous avons utilisé un ensemble de paires d'énoncés où chaque paire est composé d'un énoncé transcrit et de la meilleure sortie du système de reconnaissance qui correspond à cet énoncé.
2. Chaque énoncé utilisé est associé à une attente qui correspond à son domaine.
3. Les énoncés utilisés dans les évaluations n'ont pas été considérés pour l'écriture de la grammaire.

Les nombres des énoncés utilisés pour l'évaluation des trois modules d'analyse grammaticale de Navigator sont présentés dans le tableau suivant :

Langue	**Nombre de paires d'énoncés**
Anglais	199
Allemand	96
Italien	87

Tableau 32. Les tailles des corpus utilisés pour l'évaluation de Navigator

Ainsi, au total 762 énoncés ont été utilisés pour l'évaluation du système Navigator.

3.3.5.3 Résultats et discussion

L'unité de l'évaluation que nous avons adoptée est le slot. Ainsi, nous avons distingué entre trois types d'erreurs : insertion d'un slot, suppression d'un slot et remplacement d'un slot. Voici les résultats que nous avons obtenus[53] :

[53] Les évaluations des modules de l'allemand et de l'italien ont été faites par Marnie Lail et Valeria Lacorte respectivement.

Langue	Enoncés transcrits		Sorite de reconnaissance	
	Rappel	Précision	Rappel	Précision
Anglais	97,95	96,96	72,44	58,19
Allemand	95,86	94,74	80,32	71,69
Italien	94,43	96,74	78,27	66,23

Tableau 33. Résultats des trois versions du système Navigator

Comme nous pouvons remarquer dans le tableau précédent, les résultats sur les énoncés transcrits sont généralement assez satisfaisants tant pour le rappel que pour la précision. Nous remarquons aussi une baisse assez significative des performances avec les sorties des systèmes de reconnaissances. En effet, plus de 90% des erreurs avec la sortie du système de reconnaissance étaient dues à des remplacements/suppressions de noms propres qui ne peuvent pas être corrigés par le module de compréhension : les énoncés qui n'ont pas été traités correctement à cause d'une erreur de reconnaissance de ce genre ont une forme parfaitement grammaticale. Par exemple, lorsque le système de reconnaissance produit *I would like to go to Fondo* plutôt que *I would like to go to Trento*[54] le module d'analyse grammaticale ne peut pas corriger l'erreur de reconnaissance étant donné que l'énoncé produit est parfait tant syntaxiquement que sémantiquement.

En ce qui concerne les ambiguïtés liées au domaine de dialogue (observées avec les énoncés qui sont pertinents pour plusieurs domaines en même temps) comme : *in Trento* qui peut être un énoncé du domaine information routière, information sur les points d'intérêt et réservation hôtelière. Dans nos corpus de test pour l'anglais, l'allemand et l'italien nous avons observé respectivement : 9, 5 et 6 cas. Tous ces cas ont été correctement traités avec les énoncés transcrits grâce aux attentes du GD. Par contre, trois cas n'ont pas été traités correctement avec les sorties des modules de reconnaissance (deux cas en anglais et un cas en italien). La raison de l'échec étant l'omission des mots propres qui sont la clé principale pour la détection du thème de l'énoncé.

La différence entre les résultats des trois langues est principalement due à des raisons liées à la différence des tailles des corpus utilisés pour l'écriture des grammaires, les systèmes de reconnaissance utilisés ainsi que les différences inhérentes aux langues elles-mêmes.

Malgré la difficulté de comparer les résultats du système Oasis à ceux de Navigator (à cause des différences des langues utilisées, des unités linguistiques utilisées pour le test (arbres élémentaires vs. slots sémantiques) ainsi que les systèmes de reconnaissances utilisés), nous pouvons estimer que les résultats confirment

[54] *Trento* et *Fondo* sont deux villes italienne.

globalement ce que nous avons observé avec Oasis. Par ailleurs, cette évaluation nous a permis de montrer l'adaptation de la Sm-TAG au traitement des énoncés oraux dans le contexte de dialogues multi-domaine.

3.3.6 Discussion de la portabilité de la Sm-TAG à la lumière du système Navigator

Comme nous avons vu, la Sm-TAG est un formalisme qui combine des connaissances liées au domaine à des connaissances linguistiques indépendantes de l'application. Bien qu'elle soit l'avantage principal de la Sm-TAG, cette interaction de ces deux niveaux de connaissances peut être aussi la principale source de limitation de ce formalisme en particulier en ce qui concerne la réutilisation des ressources construites pour une application donnée à d'autres applications. Etant donné que Navigator intègre différents domaines de dialogue (dont les natures sont parfois complètement différentes), nous avons jugé bon de discuter la portabilité de la Sm-TAG à la lumière de notre expérience avec ce système. Pour ce faire, nous allons distinguer entre les trois principaux niveaux de représentation dans la Sm-TAG (les arbres lexicaux, les arbres locaux et globaux et les règles d'inférence) ainsi que deux types d'indépendance de la tâche : indépendance syntaxique et indépendance sémantique.

1. **Les arbres lexicaux :** comme nous avons vu les arbres lexicaux sont les arbres les plus nombreux dans la Sm-TAG (par exemple, dans notre grammaire de l'anglais les arbres lexicaux constituent plus de 97% du nombre total des arbres de la grammaire). Par ailleurs, la couverture lexicale est un problème central dans la construction d'un module d'analyse grammaticale quelle que soit l'approche utilisée pour ce module.
   i. Raisons grammaticales : il s'agit des arbres ancrés par des items lexicaux qui correspondent à des mots grammaticaux (prépositions, pronoms, adverbes, déterminants. Ces mots peuvent être utilisés dans tout type d'applications possible étant donné qu'ils sont nécessaires à la construction des constituants de base qui peuvent être analysés par n'importe quelle grammaire.
   ii. Raisons sémantiques : il s'agit d'arbres lexicaux qui à cause de leur nature sémantique générique peuvent être utilisés dans différentes applications (pas nécessairement toutes les applications possibles. Les nombres et les jours de la semaine sont des exemples que nous pouvons donner pour ce genre d'arbres.

   Par ailleurs, le seul cas d'indépendance de la langue était les noms des locations géographiques que nous avons utilisés pour les grammaires des trois langues.

   Ainsi, nous pouvons dire que sur le plan lexical, la Sm-TAG n'est pas fondamentalement différente des autres formalismes grammaticaux puisque la dépendance du lexique par rapport à la tâche et à la langue est soumise à des contraintes inhérentes au lexique pas à celles de la grammaire utilisée.

2. **Les arbres locaux et globaux :** les arbres locaux et globaux constituent le lieu de rencontre entre les connaissances grammaticales et les connaissances sur le domaine. A cause de cette interaction directe il

n'existe pas d'arbres qui sont indépendants du domaine à cause de raisons syntaxiques (comme c'est le cas dans les formalismes syntaxiques classiques où toutes les unités supralexicales sont indépendantes de l'application). Ainsi, l'indépendance de l'application est limitée aux raisons sémantiques. Ces arbres correspondent à des segments dont le contenu sémantique peut être utilisé dans différentes applications. Dans le cadre de Navigator nous avons eu des arbres qui correspondent à des confirmations, négations, demande de répétition. Outre ces arbres, des fragments plus significatifs des grammaires peuvent être partagés entre différents domaines. Par exemple, les arbres utilisés pour couvrir les différentes formes des dates (qui constituent entre 30% et 40% des arbres supra-lexicaux de nos grammaires de réservation hôtelière) peuvent être utilisés dans différents domaines applicatifs qui nécessitent le traitement des dates ou des expressions temporelles.

3. **Les règles d'inférence :** par définition les règles d'inférences utilisées pour l'implantent des opérations syntaxiques d'association sont indépendantes de l'application (pour des raisons grammaticales). Par contre, les règles d'inférences utilisées pour l'implantation des opérations de propagation (inductive et prédicative) sont dépendantes de la tâche. Seules les règles d'inférences qui sont associées à des arbres locaux et globaux indépendants de la tâche peuvent être portées directement.

4 Conclusion de la troisième partie

Dans cette partie, nous avons présenté deux cadres applicatifs : le premier porte sur une implantation de notre modèle des extragrammaticalités et le deuxième porte sur la réalisation de trois systèmes d'analyse linguistique du langage oral. Les objectifs de ces deux cadres étant à la fois la validation de nos études théoriques et la proposition de solutions ingénieriques permettant d'améliorer la qualité du traitement du langage oral.

4.1 Le système Corrector

Nous avons vu que l'implantation de notre modèle théorique sur les extragrammaticalités a confirmé globalement nos remarques à propos des approches précédentes. Les résultats de notre évaluation ont montré que notre approche donne des résultats légèrement supérieurs aux travaux précédents pour le traitement des répétitions et des auto-corrections et elle présente des avantages significatifs pour le traitement des faux-départs.

4.2 Analyse linguistique

4.2.1 Le système Safir

Le système Safir est un prototype que nous avons développé afin d'effectuer une première évaluation de nos idées sur l'analyse robuste du langage oral. Les résultats de l'évaluation de ce prototype nous ont permis de clarifier la portée et les limites de cette approche et ont constitué la base de nos choix pour la conception et la réalisation du système Oasis.

4.2.2 Le système Oasis

Le système Oasis combine les principales propriétés des deux systèmes précédents (Safir et Corrector) en y ajoutant de nouveaux traits que nous avons jugés nécessaires sur la base de nos expériences avec ces systèmes ou selon les requis applicatifs propres au système Oasis. Ainsi les principales propriétés de ce système peuvent être résumées dans les trois points suivants :

- Le noyau principal du système Oasis est le module d'analyse basé sur le formalisme Sm-TAG.
- L'utilisation d'une stratégie d'analyse partielle et sélective pour éviter les problèmes de sous-génération ainsi que certaines formes d'extragrammaticalités.

- L'intégration d'une stratégie de traitement des extragrammaticalités basée sur la combinaison d'un module de prétraitement et d'un module de post-traitement qui sont inspirés de notre travail sur le système Corrector.

Afin d'évaluer sa portée et ses limites, nous avons évalué le système Oasis selon deux méthodes : une méthode quantitative et une méthode qualitative.

4.2.2.1 Evaluation quantitative

Les résultats de cette évaluation ont montré que la performance de notre système est comparable à celle des autres systèmes dans la littérature. Les erreurs de reconnaissance ont été la source principale des erreurs d'analyse syntaxique. Bien que notre système était capable de donner une analyse correcte dans environ 35% des cas d'énoncés avec des erreurs de reconnaissance, nous estimons que l'amélioration de l'interaction avec le module de reconnaissance permet d'augmenter la robustesse du module d'analyse linguistique vis à vis des erreurs de reconnaissance.

4.2.2.2 Evaluation qualitative

Cette évaluation a été réalisée dans le cadre d'une campagne nationale qui a regroupé les principaux laboratoires français qui travaillent sur l'analyse linguistique du langage oral. L'analyse détaillée des résultats de cette campagne nous a permis de constater la bonne performance de notre système à traiter la majorité des phénomènes grammaticaux et extragrammaticaux observés dans notre corpus. A cause de raisons inhérentes à la méthode d'évaluation par défi ainsi qu'au fait que la campagne d'évaluation est toujours en cours, il ne nous a pas été possible de comparer finement les résultats qualitatifs obtenus avec les autres systèmes impliqués dans la campagne.

4.2.3 Le système Navigator

Le système Navigator est un système de compréhension destiné au traitement des dialogues multi-domaines orientés par la tâche. Les spécificités principales de ce système sont :

- Architecture modulaire où nous avons différentes unités d'analyse qui correspondent chacune à un domaine particulier du dialogue.
- Utilisation des attentes de haut niveau pour le guidage des modules d'analyse grammaticale, d'arbitrage et d'analyse sémantique.
- Traitement des *N* meilleurs hypothèses de reconnaissance.

L'évaluation de Navigator a confirmé globalement les résultats obtenus avec Oasis (bien que Navigator soit conçu pour traiter des dialogues multi-domaines).

Conclusion et perspectives

1 Bilan général

Dans cette thèse, notre travail a été motivé par un objectif principal qui est la formalisation et la réalisation d'un système d'analyse linguistique automatique du langage oral capable de combiner la robustesse et la profondeur. En ce qui concerne notre contribution dans cette thèse, elle peut être résumée dans les points suivants :

1.1 Traitement des extragrammaticalités

Notre travail sur ce plan s'articule autour de deux axes complémentaires :

1.1.1 Analyse de corpus

Sur cet axe, nous avons mené une étude des extragrammaticalités sur la base du *Trains Corpus* de l'université de Rochester qui contient 93 dialogues dont nous avons extrait environ 7000 cas d'extragrammaticalités dont 928 cas d'extragrammaticalités supralexicales[55]. Les points clés de notre étude des extragrammaticalités sont résumés dans ce qui suit :

- Dans notre typologie nous avons distingué entre les extragrammaticalités lexicales et les extragrammaticalités supralexicales d'une part et d'autre part nous avons proposé la prise en considération de l'incomplétude comme une forme d'extragrammaticalité.
- Contrairement aux études précédentes qui ont proposé un schéma unique pour les extragrammaticalités (Shriberg, 1994), nous avons proposé quatre schémas correspondant chacun à un type particulier de phénomènes. Cela nous permet de refléter plus fidèlement les différentes propriétés de ces phénomènes et par conséquent adopter une approche plus adaptée pour les traiter.
- Afin d'avoir une analyse plus précise du corpus, nous avons étiqueté non seulement les extragrammaticalités mais aussi les fausses extragrammaticalités, c'est-à-dire, les cas normaux qui peuvent être pris pour une extragrammaticalité pour une raison ou une autre.

Cette typologie nous a permis d'une part de constater que les extragrammaticalités sont des phénomènes qui présentent une régularité assez importante et d'autre part, cela nous a permis de mettre la lumière sur les raisons de l'échec des approches syntaxiques des travaux antérieurs. Par ailleurs, nous avons pu constater à travers l'analyse des principaux cas observés que la production des extragrammaticalités est intimement liée à la grammaire de la langue dans laquelle ils sont produits.

[55] Les extragrammaticalités supralexicales couvrent les répétitions, les autocorrections, les faux-départs et les incomplétudes.

1.1.2 Réalisation du système Corrector pour le traitement des extragrammaticalités

Ce système est basé sur la combinaison de la reconnaissance de patrons et de l'analyse partielle. Il augmente la robustesse du système vis à vis des extragrammaticalités et il permet d'affiner l'analyse en évitant les erreurs d'interprétation qui peuvent être causées par les autocorrections, les faux-départs, etc. Les avantages de notre approche se résument dans les points suivants :

- Le système a été conçu pour opérer comme une phase de pré-traitement au sein de systèmes d'analyse linguistique et de dialogues plus larges. Ainsi, le système a été conçu pour être facilement portable d'une application à une autre (voire d'une langue à une autre) puisqu'il est complètement indépendant des composantes du système au sein duquel il s'intègre.
- Notre approche qui combine les techniques de reconnaissance de patrons et d'analyse superficielle nous a permis d'optimiser le rapport simplicité/efficacité pour le traitement des différents phénomènes. En particulier, cela nous a permis de prendre en considération un contexte plus large (que celui utilisé dans les approches à base de N-grams). Par ailleurs, à notre connaissance, notre étude est la première qui utilise la syntaxe seulement pour la détection des faux-départs.
- Intégration d'informations de haut niveau dans la détection des extragrammaticalités notamment à l'aide de grammaires sémantique.
- Implantation de règles et de patrons de contrôle (basés sur les modèles des fausses extragrammaticalités) afin de réduire les surgénérations du système.
- L'évaluation de notre système sur 581 cas d'extragrammaticalités dont 309 cas d'extragrammaticalités supralexicales a montré une amélioration dans les taux de détection et de délimitation des différents phénomènes considérés par rapport aux travaux précédents.
- Nous avons porté les patrons obtenus pour l'anglais pour le traitement des extragrammaticalités en français. L'évaluation du module français a montré son efficacité pour le traitement des répétitions et des autocorrections.

1.2 Analyse grammaticale

1.2.1 La Grammaire Sémantique de Substitution d'Arbres (S-TSG)[56]

La Grammaire Sémantique de Substitution d'Arbres est une formalisation que nous proposons pour une grammaire sémantique, approche couramment utilisée dans le domaine de l'analyse linguistique du langage oral. Deux avantages distinguent la S-TSG d'une grammaire sémantique classique :

[56] Semantic Tree Substitution Grammar.

- **Avantages théoriques :** la S-TSG est un formalisme bien défini mathématiquement et dont les propriétés linguistiques sont assez claires (notamment en ce qui concerne le lien entre le lexique et la grammaire d'une part et la syntaxe et la sémantique d'autre part). Cela permet d'établir des comparaisons rigoureuses entre ce formalisme et les autres formalismes existants d'une part et d'autre part, cela permet de clarifier la portée et les limites de ce formalisme par rapport au traitement du langage oral.
- **Avantages pratiques :** comparée à la grammaire sémantique classique, la S-TSG se distingue par une structure hiérarchisée des sources d'information selon trois niveaux : arbres lexicaux, arbres locaux et arbres globaux. Cela rend l'écriture et la modification de la grammaire une tâche plus facile.

1.2.2 La Grammaire Sémantique d'Association d'Arbres (Sm-TAG)[57]

A notre connaissance, la Sm-TAG est le premier formalisme grammatical conçu spécifiquement pour le langage oral. Sa particularité principale est l'intégration de connaissances extralinguistiques (modèle sémantique de la tâche) dans la représentation des connaissances linguistiques.

D'un point de vue de traitement, la Sm-TAG est un compromis entre les grammaires sémantiques (approches robustes mais trop superficielles) et les grammaires syntaxiques classiques (approches fines mais peu robustes). Dans le contexte de la problématique de la thèse, ce formalisme contribue à deux niveaux :

- Il augmente la robustesse puisque, d'une part, il est conçu sur la base de la syntaxe de l'oral et d'autre part, il prend en considération les informations sur la tâche qui sont une source assez fiable dans le contexte de dialogues orientés par la tâche.
- Sur le plan de la profondeur, nous avons montré que nous pouvons analyser avec la Sm-TAG les principaux les phénomènes syntaxiques que nous pouvons analyser avec les formalismes classiques.

1.2.3 Systèmes d'analyse grammaticale

Deux systèmes ont été construits pour valider les formalismes S-TSG et Sm-TAG :

4.2.3.1 Le système Safir

Le système Safir est un prototype que nous avons réalisé afin de faire une première évaluation de notre approche. Ce prototype a été réalisé avec une grammaire de type S-TSG convertie en un réseau de transition récursif enrichi par les traits suivants :

- Une stratégie sélective par grammaire de nettoyage qui permet d'ignorer les parties du message que le système ne peut pas traiter.

[57] Semantic Tree Association Grammar.

- Une approche d'analyse partielle : cette approche consiste à relaxer les contraintes d'analyse dans les cas où l'on ne peut pas obtenir un arbre d'analyse dont la racine est l'axiome de la grammaire. Cela permet à des unités de rang inférieur d'être considérées comme des unités bien formées même si elles sont complètement indépendantes du reste des unités.

4.2.3.2 Le système OASIS

Les composantes principales du système OASIS sont les suivantes :

1. Un module de pré-traitement basé sur un ensemble de patrons portés de l'anglais. Ce module a pour fonction de normaliser les répétitions et les autocorrections.
2. Un module d'analyse robuste basé sur le formalisme Sm-TAG.

Le système Oasis a été évalué selon deux méthodes :

2. **Une évaluation classique :** cette évaluation a été faite avec 210 énoncés (non utilisés pour l'écriture de la grammaire). Afin de tester l'adaptation de notre approche au traitement des erreurs de reconnaissance, nous avons lu et enregistré ces énoncés. Nous avons ensuite analysé les fichiers obtenus avec le système de reconnaissance Raphaël. Les résultats ont montré l'adaptation de notre approche au traitement des énoncés avec des erreurs de reconnaissance ainsi qu'un niveau de couverture lexical et sémantique acceptable.
3. **Evaluation quantitative :** cette évaluation est basée sur la méthode d'évaluation par défi qui est une version modifiée de la méthode DCR (Antoine *et al.*, 2001). Cette évaluation s'est déroulée en collaboration avec sept collègues de quatre laboratoires français dans le cadre d'une campagne menée par le GT "Compréhension robuste de la langue" du GDR-I3. Les résultats de notre système étaient satisfaisants pour les quinze principaux phénomènes observés dans notre corpus de test.

4.2.3.3 Le système Navigator

Le système Navigator est un système de compréhension destiné à traiter des dialogues multi-domaine en anglais, allemand et italien. La différence principale entre Navigator et Oasis est l'adoption d'une approche modulaire pour le traitement des énoncés dans le contexte de dialogues multi-domaine. Ainsi, au lieu d'avoir une grammaire pour tous les domaines de dialogue, nous avons différentes parties de la grammaire réparties sur des unités indépendantes et qui partagent un ensemble d'arbres (lexicaux, locaux et globaux) qui sont indépendant du domaine. L'activation de chacune des unités se fait sur la base des attentes fournies par le gestionnaire de dialogue. L'évaluation de ce système a confirmé globalement les résultats que nous avons obtenus avec Oasis. Par ailleurs, ces résultats ont montré l'adaptation de la Sm-TAG au traitement des dialogues multi-domaine.

2 Perspectives à court-terme

Notre travail en cours se focalise sur trois axes :

1. Réalisation du module de compréhension dans le cadre du projet européen NICE[58] : les deux principaux défis pour la compréhension dans le cadre de ce projet sont :
   i. Le dialogue est orienté par le domaine et non par la tâche. Cela nécessite la création d'un module de compréhension qui est capable de traiter des thèmes assez variés qui peuvent être abordés par les utilisateurs du système de dialogue tout en respectant un degré minimal de profondeur.
   ii. L'intégration des gestes aux énoncés parlés pour la compréhension de l'entrée multi-modale.

   Ce projet étant encore dans sa première année, notre réalisation se limite actuellement à la proposition de l'architecture du module de compréhension qui est une extension de celle de Navigator. En effet, nous avons conçu une architecture parallèle qui combine un module à base de Sm-TAG similaire à Oasis à un module de détection de thème.
2. La campagne d'évaluation par défi : nos résultats ont montré que la méthode d'évaluation par défi est bien adaptée au diagnostic des différentes propriétés d'un système d'analyse linguistique automatique du langage oral. Cette méthode n'est cependant pas parfaite notamment en ce qui concerne les possibilités de comparaison des résultats obtenus par les différents systèmes impliqués dans la campagne. Ainsi, nous sommes en train d'explorer avec les collègues impliqués dans cette campagne, l'homogénéisation des critères de test afin de pouvoir comparer objectivement les différents systèmes d'une part et d'autre part pour pouvoir aller plus loin dans le diagnostique des raisons d'échec et de réussite de chaque système.
3. La méthode DCR étendue : après nos tests prometteurs sur trois phénomènes syntaxiques de la méthode DCR étendue que nous avons proposée, nous sommes en train de travailler sur la généralisation de cette méthode à l'évaluation du reste des phénomènes syntaxiques ainsi qu'aux phénomènes sémantiques et pragmatiques. Cela permettra d'utiliser cette méthode pour évaluer non seulement l'analyse linguistique (comme c'est le cas avec la version actuelle de la méthode) mais aussi la compréhension. Autrement dit, cela permettra la prise en considération de l'historique de l'interaction pour évaluer la qualité de l'analyse sémantique produite par le système.

[58] Le projet NICE est aussi un projet de trois ans. Les principaux partenaires impliqués dans ce projet sont Telia (Suède), LiquidMedia (Suède), Philips (Allemagne), LIMSI (France) et NisLab (Danemark). L'objectif principal de ce projet est la construction d'un système de dialogue multi-modal avec des agents virtuels qui représentent des personnages issus des contes de Hans-Christian Anderson.

3 Perspectives à plus long terme

3.1 Modélisation des extragrammaticalités

Malgré la confirmation de notre modèle théorique par son application dans le cadre du système Corrector, il nous semble que ce modèle peut être enrichi sur deux plans :

- **Intégration de la prosodie :** comme nous avons vu dans l'état de l'art, les travaux précédents ont mis en évidence l'intérêt de la prosodie pour le traitement des extragrammaticalités notamment en ce qui concerne la détection de ces phénomènes. Ainsi, nous estimons que l'intégration des informations prosodiques à notre modèle, essentiellement basé sur la syntaxe, permettra d'augmenter sa couverture.
- **Typologie syntaxique plus fine des extragrammaticalités :** la place des extragrammaticalités au sein d'une théorie syntaxique générale de l'oral reste un objet à discussion. Pour donner une base à une réponse scientifique à cette question, une typologie linguistique fine qui comprend à la fois le niveau grammatical et les niveaux discursif et pragmatique nous semble une démarche indispensable. Cela permettra par ailleurs de créer un cadre général qui englobe à la fois les extragrammaticalité et la Sm-TAG.

3.2 La Sm-TAG

Vu les avantages de la Sm-TAG à intégrer des sources de connaissances diverses dans le même cadre, il nous semble utile d'explorer l'intégration des connaissances issues de modalités différentes dans le cadre d'une version multi-modale de la Sm-TAG. Cela permettra, en particulier, de désambiguïser la référence des déictiques dans le contexte d'interaction multi-modale (comme celui du projet NICE).

Bibliographie

1. Références bibliographiques[59]

ABEILLE, A., GODARD, Danièle, (1997), The syntax of french negative adverbs, in Anne AVBEILLE, Danièle GODARD, Philip MILLER, *The major syntactic structures of French*, ESSLLI Summerschool, Aix-en-Provence.

ABEILLE, Anne, (1993), *Les nouvelles syntaxes :* Grammaires d'unification et analyse du français, Paris : Armand Colin.

ABNEY, Steven, (1991), Parsing by chunks, in Robert BREWICK, Steven ABNEY, Carol TENNY (éditeurs), *Principle-based parsing*, Dordrecht : Kluwer Academic Publishers.

ABNEY, Steven, (1995), Partial Parsing via Finite-State Cascades, *Journal of Natural Language Engineering*, **2**(4): 337-344.

ABNEY, Steven, (1996), Chunk Style book, Working draft, University of Tübengen, http://www.sfs.nphil.uni-tuebingen.de/~abney/Papers.html#96i.

AÏT-MOKHTAR, Salah, CHANOD, Jean-Pierre, (1997), Incremental finite-state parsing, In Processing of the 5th International conference on Applied Natural Language Processing (ANLP), Washington DC, USA, p. 72-79.

AÏT-MOKHTAR, Salah, CHANOD, Jean-Pierre, (1997), Incremental finite-state parsing, In Proceedings of Applied Natural Language Processing, Washington, DC. April.

ALSHAWI, Hiyan (éditeur), (1992), *The Core Language Engine*, Cambridge, Massachusetts: MIT Press.

ALTMANN, G.T.M, (1999), Thematic role assignment in context, *Journal of Memory and Language*, 41, 124-145.

ANDROWS, A., (1985), The major functions of noun phrase, in Timothy Shopen (éditeur), *Language typology and syntactic description*, Vol. 1, Cambridge : Cambridge University Press.

ANTOINE, Jean-Yves, (1994), *Coopération-syntaxe sémantique pour la compréhension de la parole spontanée*, Thèse Signal Image Parole, Institut National Polytechnique de Grenoble.

ANTOINE, Jean-Yves, ZEILIGER, Jérôme, CAELEN, Jean., (1998), DQR Test suites for a qualitative evaluation of spoken dialog systems : from speech understanding to dialog strategy, Proceedings of the International Conference on Language Ressources and Evaluation LREC'98, Grenade, Espagne.

ANTOINE, Jean-Yves, BOUSQUET-VERNHETTES, Caroline, GOULIAN, Jérôme, KURDI, Mohamed-Zakaria, Rosset, Sophie, VIGOUROUX, Nadine, VILLANEAU, Jeanne, (2002), Predictive and objective evaluation of speech understanding: the "challenge" evaluation campaign of the I3 speech workgroup of the French CNRS, Third International Conference on Language Ressources and Evaluation LREC02, Las Palmas.

ANTOINE, Jean-Yves, GOULIAN, Jérôme, (2001), *Word order variations and spoken man-machine dialogue in French : a corpus analysis on the ATIS domain*, Corpus Linguistics'2001, Lancaster, UK.

ANTOINE, Jean-Yves, SIROUX, Jacques, CAELEN, Jean, VILLANEAU, Jeanne, GOULIAN, Jerome, AHFHAF, Mohammed, (2000), Obtaining predictive results with an objective evaluation of spoken dialogue systems : experiments with the DCR assessment paradigm, Proceedings of the International Conference on Language Ressources and Evaluation LREC'2000, Athènes, Grèce.

[59] Cette liste contient uniquement les références citées dans la thèse.

ARLAN, David, SHAW, Mary, (1993), An Introduction to Software Architecture, In V. Ambriola and G. Tortora (editeurs), Advances in Software Engineering and Knowledge Engineering, Series on Software Engineering and Knowledge Engineering, Vol 2, Singapore: World Scientific Publishing Company, pp. 1-39, 1993. http://www-2.cs.cmu.edu/afs/cs/project/able/www/paper_abstracts/intro_softarch.html

ASHBY, W., (1981), The loss of the negative particle *ne* in French : a syntactic change in progress, Language, 57, p.674-687.

BATES, E., CARNEVALE, G. F., (1997), New directions in research on language development, rapport technique university of California San Diego. http://www.ecs.soton.ac.uk/~harnad/Papers/Py104/bates-carnevale.html

BATLINER, Anton, *et al.*, (2000), The prosody module, *in* Wolfgang Wahlster (éditeur), *Verbmobil:* foundations of speech-to-speech translation, Berlin : Springer.

BEAR, J., DOWDING, J., SHRIBERG, E., (1992), Integrating multiple knowledge sources for detection and correction of repairs in human-computer dialogs, In Proceedings of the 30th Meeting of the Association for Computational Linguistics, (Newark, Delaware), June.

BERNSEN, N. O., DYBKJÆR, L., (1999), A theory of speech in multimodal systems, Proceedings of the ESCA Tutorial and Research Workshop on Interactive Dialogue in Multi-Modal Systems, Irsee, Germany, June 1999.

BERNSEN, Niels-Ole, (Editeur), (2002), Progress Report on the Natural language understanding, Dialogue Management, Response Generation, and Speech Synthethsis components, Vico deliverable D11, NISLab, November.

BESSAC, M., CAELEN, G., (1995), Analyses pragmatiques, prosodiques et lexicales d'un corpus de dialogue oral homme-machine, JADT'95, Roma, Italie, 363 :370.

BLANCHE-BENVENISTE, Claire, (1990), *le français parlé : études grammaticales*, Editions du CNRS, Paris.

BLANCHE-BENVENISTE, Claire, (1997), *Approches de la langue parlée en français*, Ophrys, Paris.

BLANCHE-BENVENISTE, Claire, JEANJEAN, Colette, (1987), *Le Français parlé* : transcription et édition, Didier Erudition.

BOD, R., (1995), *Enriching Linguistics with Statistics :* Performance Models of Natural Language, Ph.D. Dissertation, University of Amsterdam, Holland.

BOD, R., KAPLAN, R., (1998), A Probabilistic Corpus-Driven Model for Lexical Functional Analysis, Proceedings COLING-ACL-98, Montreal, Canada.

BOLAND, J. D., TANENHAUS, M. K., GARNSEY, S. M., CARL-SON, G. N., (1995), Verb argument structure in parsing and interpretation: Evidence from wh-questions, *Journal of Memory and Language*, 34, 774–806.

BONNEMA, R., *et al.*, (1999), Evaluation results NLP component OVIS2, 31 mei, Amsterdam/ Eindhoven/ Groningen. http://odur.let.rug.nl:4321/publijst.html

BOUFADEN, Narjes, (1998), *Analyse syntaxique robuste des textes de dialogues orax*, Mémoire de maîtrise, Université Laval, Québec.

BOULAKIA, G., (1983), Phonosyntaxe du français, Revue Internationale du Traitement Automatique du Langage, vol. 24.

BOULAKIA, G., (1987), Ambiguïté et intonation, in C. FUCHS (éditeur), *L'Ambiguïté et la Paraphrase :* opérations linguistiques, processus cognitifs et traitements automatisés, Caen : Centre de Publications de l'Université de Caen.

Bousquet-Vernhettes, Caroline, (2002), *Compréhension robuste de la parole spontanée dans le dialogue oral homme-machine* – Décodage conceptuel stochastique, Thèse de doctoral de l'Université Paul SABATIER-Toulouse-III.
http://www.irit.fr/recherches/MODEL/DIAM/Membres/page.php3?name=bousquet

BRANIGAN, H. P., PICKERING, M. J., LIVERSEDGE, S. P., STEWART, A. P., URBACH, T. P., (1995), Syntactic priming: Investigating the mental representation of language, Journal of Psycholinguistic Research, 24, 489–506.

BRILL, Eric, (1993), *A corpus based approach to language learning*, Ph.D. Dissertation, University of Pennsylvania, 1993.

BRISCOE, Ted, (1987), Deterministic parsing and unbounded dependencies, Third Conference of the European Chapter of the Association for Computational Linguistics, University of Copenhagen, Denmark, 1-3 Apr. 1987, p. 211-217.

BRISCOE, Ted, CAROLL, Jhon, (1993), Generalized probabilistic LR parsing of natural language with unification grammars, Computational linguistics 19(1) : 25-59.

BROWN, John, BURTON, Richard, (1975), Multiple representations of knowledge for tutorial reasoning, In D. BOBROW, A. COLLINS, (éditeurs), *Representation and understanding :* studies in cognitive science, New York : Academic press.

BURTON, Richard, (1976), Semantic grammar : an engineering technique for constructing natural language understanding systems, BBN report 3453.

CAELEN, J., *et al*, (1990), Architecture et fonctionnement du système DIRA : de l'acoustique aux niveaux d'analyse linguistique, Traitement du signal 4,Vol. 7.

CAELEN, Jean, *et al.*, (1997), Les corpus pour l'évaluatin du dialogue homme-machine, ARC B2, Journée JST-FRANCIL, Avignon.

CAILLAUD, Bertrand, (1996), *Apprentissage de connaissances prosodiques pour la reconnaissance automatique de la parole*, Thèse Signal Image Parole, l'Institut National Polytechnique de Grenoble.

CAPELLE, Guy, FREROT, Jean-Louis, *Grammaire de base*, Paris : Hachette.

CARABALLO, Sharon, CHARNIAK, Eugene, (1996), Figures of Merit for Best-First Probabilistic Chart Parsing, Brown University, Rapport Technique CS-96-12, April.

CARBONEL, N., PIERREL, JM., (1986), Architecture and knowledge sources in human computer oral dialogue system, Workshop OTAN, Corse, France.

CARBONELL, J. G., HAYES, P.J., (1984), Recovery strategies for parsing

CARBONNEL, J.G., HAYES, P.J., (1983), Recovery strategies for parsing extragrammatical language, American Journal of Computational linguistics, 9(3-4), pp123-146.

CARPENTER, P. A., JUST, M. A., (1988), The role of working memory in language comprehension, in D. KLAHR, K. KOTOVSKY, (éditeurs), *Complex information processing*: the impact of Herbert A. Simon, Hillsdale, NJ : Erlbaum.

CARRE, René, DEGREMONT, Jean-François, GROSS, Maurice, PIERREL, Jean-Marie, SABAH, Gérard, (1991), *Langage humain et machine*, Paris : Presses du CNRS.

CAVAZZA, Marc, (1998a), An integrated parser for TFG with explicit tree typing, Proceedings of the International TAG Workshop, Philadelphia, 28-31July.

CAVAZZA, Marc, (1998b), Synchronous TFG for speech translation, Proceedings of the International TAG Workshop, Philadelphia, 28-31July.

CHAFE, W. L., (1975), Giveness, contrastiveness, definiteness, subjects, topics and point of view, in C. N. LI (éditeur), Subject and topic, New York, Academic Press, pp. 25-55.

CHELBA, C., ENGLE D., JELINEK, F., JIMENEZ, V., KHUDANPUR, S., MANGU, PRINTZ, H., RISTAD, E., ROSENFELD, R., STOLCKE, A., WU, D., (1997), Structure and Performance of a Dependency Language Model, Proceedings of EUROSPEECH, 2775-2778, Rhodes, 22 – 25 September, Greece.

COCCARO, N., JURAFSKY, D., (1998), Towards better integration of semantics predictors in statistical language modeling, ICSLP'98, Sydney, Australia.

COLE, R. A. (éditeur), (1996), Survey of the state of the art in human language technology, National Science Foundation (USA). http://cslu.cse.ogi.edu/HLTsurvey/

COLINEAU, Nathalie, (1997), *Etude des marques discursifs dans le dialogue finalisé*, Thèse de l'université Joseph Fourier-Grenoble I.

CORAZZA, Anna, (1999), An inter-domain portable approach to Interchange Format construction, Eurosepeech99, Budapest, Hungary, September 6-9, pp. 2419-2422.

CORBLIN, Francis, (1995), Compositionnality and complexity in multiple negation, Interest groupe in pure and applied logic Bulletin 3:3-2.449-473.

CORE, Marc G., (1999), *Dialog parsing :* from speech repairs to speech acts, Ph.D. University of Rochester, USA.

CORE, Mark, SCHUBERT, Lenhart, (1998), Implementing Parser Metarules that Handle Speech Repairs and Other Disruptions, Presented at the 11th International FLAIRS Conference, Sanibel Island, FL., May.

CRAIN, S., STEEDMAN, M., (1985), On not being led up by the garden path: the use of context by the psychological syntax processor, in D. R. Dowty *et al.*, (éditeurs), *Natural Language Parsing:* computationl and theoretical perspectives, Cambridge: Cambridge University Press.

CROCKER, M.W., (1999), Mechanisms for sentence processing, In Garrod & Pickering (éditeurs), *Language Processing*, London : Psychology Press.

CROOKER, M. W., (1996), Mechanisms for sentence processing, The university of Edinburgh, Centre for Cognitive Science, Research parper EUCCS/RP-70, November.

CUTLER, A., (1996), Prosody and the word boundary problem, in J., MORGAN, *et al*, (sous la direction), *Signal to Syntax :* bootstrapping from speech to grammar in early acquisition, Mahwah : Laurence Erlbaum associates.

DE MORI, R., (1994), Apprentissage automatique pour l'interprétation sémantique, XXèmes Journée d'Etude sur la Parole (JEP'94), Trégastel, Juin.

DE SAUSSURE, (1968), *Cours de linguistique générale*, édition critique par R. ENGLER, Otto Wiesbaden : HArrassowitz. Reproduction de l'édition originale, 1989.

DE SMEDT, K., KEMPEN, G., (1990), Segment Grammar a formalisme for incremental generation, In C. PARIS *et al.*, (éditeurs), *Natural language generation and computational linguistics*, Dordrecht : Kluwer Academic Publisher.

DE MORI, Renato, (1994), Apprentissage automatique pour l'interprétation sémantique, Trégastel, XXèmes JEP, Juin.

DELMONTE, Rodolfo, BIANCHI, Dario, (2002), From deep to partial understanding with GETARUNS, .2nd Workshop on Robust Methods in Analysis of Natural language Data Romand2002, Frascati-Rome, July 17.

DOWDING, J., GAWRON, J. M., APPELT, D., BEAR, J., CHERNY, L., MOORE, R., and MORAN, D., (1993), GEMINI : A natural language system for spoken-language understanding, In Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics, Columbus, Ohio.

DUBOIS, Jean, *et al.*, (1994), *Dictionnaire de linguistique et des sciences du langage*, Paris : Larousse.

ERMAN, L. D., *et al.*, (1980), HEARSAY-II speech understanding system : integrating knowledge to resolve uncertainty, Comput. Surv. 12, 213-253.

FARGUES, J., (1989), Graphes conceptuels et compréhension du langage naturel, Actes du congrès Sémantica, EC2, Paris, juin.

FINKLER, Wolfgang, (1997), A Descriptive View of Human Self-Corrections as the Basis of a Constructive Approach to Automatic Self-Corrections during Incremental Generation, Computational Psycholinguistics'97, August 10-12, San Francisco.

FLICKINGER, Dan, *et al.*, (2000), HPSG analysis of English, *in* Wolfgang Wahlster (éditeur), *Verbmobil: foundations of speech-to-speech translation*, Berlin : Springer.

FOX TREE, J.E., SCHROCK, J.C., (1999), Discourse markers in spontaneous speech: Oh what a difference an "oh" makes, *Journal of Memory and Language*, 40, 280-295.

FREUD, Sigmund, (1901), Psychopathology of Everyday Life, traduit en anglais par A. A., BRILL (1914), Classics in the History of Psychology, Toronto Ontario : York University electronic resources. http://psychclassics.yorku.ca/Freud/Psycho/

FRIEDMAN, N., GOLDSZMIDT, M., (1996), Learning bayesien Network with local structure, In proceedings of twelfth conference on Uncertainty in Artificial Intelligence (UAI). http://www.cs.huji.ac.il/~nir/abstracts/FrG1.html

FUCHS, Catherine, (1996), *Les ambiguïtés du français*, Paris : Ophrys.

GADET, F., (1992), *Le français populaire*, Paris : Armand Colin, 1992.

GADET, François, (1989), *Le français ordinaire*, Paris : Armond Colin.

GARETT, M.F., (1988), Processes in language production, *in* F. J. NEWMEYER (éditeur), *Linguistics* : the Cambridge Survey, vol III: language: psychological and biological aspects, Cambridge: Cambridge University Press.

GAVALDÀ, Marsal, (2000), *Growing semantic grammar*, Ph.D. Dissertation, Carnegie Mellon University.

GEE, J. P., GROSJEAN, F., (1983), Performance structures : A psycholinguistic and linguistic appraisal, *Cognitive Psychology*, 15 :411-458.

GILBERS, Dicky, DE HOOP, Helen, (1998), Conflicting constraints: an introduction to optimality theory, *Lingua* 104, 1-12. www.folli.uva.nl/CD/1999/library/pdf/GILB-DHO.pdf

GIMENEZ, Jean-Marc, (2000), L'aspirateur de sites, précieux pour la veille en ligne, 01 informatique no 1589, 19 Mai.

GÖDEL, Kurt, (1931), On formally undecidable propositions of principia mathematica and related systems http://www.ddc.net/ygg/etext/godel/godel3.htm

GODFREY, J.J., HOLLIMAN, E.C., McDANIEl, J., (1992), SWITCH-BOARD: Telephone speech corpus for research and develop-ment, Proc. IEEE Internat. Conf. Acoust. Speech Signal Pro-cess,pp. 517.520.

GOLDSMITH, John, (1990), *Autosegmental and metrical phonology*, Cambridge, Massachusetts: Blackwell.

GOODMAN, Joshua, (1999), Semiring Parsing, *Computational linguistics*, Volume 25, Number 4, pp 573 – 605.

GOULIAN, Jérôme, ANTOINE, Jean-Yves, POIRIER, Franck, (2002), Compréhension automatique de la parole et TAL : une approche syntaxico-sémantique pour le traitement des inattendus structuraux du français parlé, proc. TALN'2002. Nancy, France. Juin 2002.

GREFENSTETTE, Gregory, (1999), Light Parsing as Finite-State Filtering, In Andras, Kornai, (éditeur), *Extended Finite State Models of Language*, Cambridge : Cambridge University Press, 1999.

GRINBERG, D., LAFFERTY, J., SLEATOR, D., (1995), A robust parsing algorithm for link grammars, (1995), Proceedings of the Fourth International Workshop on Parsing Technologies, Prague and Karlovy Vary, September 1995, pp. 111-125. Also issued as technical report CMU-CS-95-125, Department of Computer Science, Carnegie Mellon University.

GROSJEAN, F., Hirt, C., (1996), Using prosody to predict the end of sentences in English and French: Normal and brain-damaged subjects, *Language and Cognitive Processes*, 11 (1/2), 107-134.

GROSZ, B., JOSHI, A., Weinstein, S., (1995), Centering : a framework for modeling the local coherence of discourse, Computational linguistics, 21 : 203-225.

GRUNE, Dick, JACOBS, Ceriel, (1990), *Parsing techniques* : a practical guide, Chichester: Ellis Horwood. http://www.cs.vu.nl/~dick/PTAPG.html.

HALBER, A., (1999), *Stratégie d'analyse pour la compréhension de la parole :* vers une approche à base de grammaires d'arbres adjoints lexicalisés, thèse de L'école Nationale Supérieure des Télécommunications, ENST, Paris.

HARDT, Daniel, (1997), An Empirical Approach to VP Ellipsis, *Computational Linguistics*, 23(4).

HARTSUIKER, R.J., KOLK, H.H.J, (1998), Syntactic facilitation in agrammatic sentence production, *Brain and Language, 62* (2), 221-254.

HEEMAN, Peter A., (1997), *Speech Repairs, Intonational Boundaries and Discourse Markers:* Modeling Speakers' Utterances in Spoken Dialog, Ph.D. Dissertation, University of Rochester.

HEEMAN, Peter, ALLEN, James, (1994), Detecting and Correcting Speech Repairs, In Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics (ACL-94), pages 295-302, Las Cruces, June.

HEEMAN, Peter, ALLEN, James, (1995), The train 93 dialogs, TRAINS Technical note94-2, The University of Rochester computer science department, March.

HEMPHILL, C., GODFREY, J., DODDINGTON, G., (1990), The ATIS Spoken Language Systems pilot corpus, Proc. DARPA Speech and Natural Language Workshop, Hidden Valley, Pa., pp. 96-101.

HEMPHILL, *et al*, (1990), The ATIS Spoken Language Corpus, Workshop on Speech and Natural Language, Hidden Valley (PA), June.

HINDLE, D., (1983), Deterministic parsing of syntactic non-fluencies, Proceedings of the 21 meeting of the association of computational linguistics, pp 123-128.

HINRICHS, E. W., *et al.*, (2000), Robust chunk parsing for spontaneous speech, *in* Wolfgang Wahlster (éditeur), *Verbmobil :* foundations of speech-to-speech translation, Berlin: Springer.

HOBBS, Jerry R., (1978), Resolving pronoun references, Lingua 44, 339-352.

HOCKETT, C. F., (1967), Where the tongue slips, there slip I, In *To honor Roman Jacobson*: Vol2, The Hague : Mouton.

HOFMANN, Thomas, PUZICHA, Jan, (1999), Latent Class Models for Collaborative Filtering, in proceedings of IJCAI'99, July 31- August 6, Stockholm, Sweden.

HOLLARD, S., (1997), L'organisation des connaissances dans le dialogue orienté par la tâche, rapport technique 1-97, Geod CLIPS-IMAG, Grenoble.

http://fdlwww.kub.nl/~krahmer/index2.html#book

JELINEK, F., LAFFERTY, J.D., MERCER, R.L., (1990), Basic methods of probabilistic context free grammars, Research report RC 16374 (#72684), IBM, Yorktown Heights, New York 10598.

JOSHI, A., LEVY, L.S., TAKAHASHI, M., (1975), Tree adjunct grammar, *Journal of Computer and System Science*, vol. 21, no. 2.

JOSHI, Aravind, (1996), Parsing techniques, In COLE, R. A. (éditeur), *Survey of the state of the art in human language technology*, National Science Foundation (USA). http://cslu.cse.ogi.edu/HLTsurvey/

JOSHI, Aravind, SCHABES, Yves, (1999), Tree-Adjoining Grammars, talk at Michigan State University, March 15, http://www.cis.upenn.edu/~joshi/

JURAFSKY, Daniel, MARTIN, James H., (2000), Speech and language Processing: an introduction to speech recognition, Natural language processing, and computational linguistics, Prentice Hall Cliffs, New Jersey.

KAISER, Ed, JOHNSTON, Michael, HEEMAN, Peter A., (1999), PROFER: Predictive, Robust Finite-state Parsing for Spoken Language, Proceedings of ICASSP, Volume II, pgs. 629-32, March.

KARGER, Reinhard, WAHLSTER, Wolfgang, (2000), Facts and figures about the Verbmobil project, *in* Wolfgang Wahlster (éditeur), *Verbmobil :* foundations of speech-to-speech translation, Berlin: Springer.

KARTTUNEN**,** Lauri, (2000), Applications of Finite-State Transducers In Natural Language Processing, In Proceedings of CIAA-2000. Lecture Notes in Computer Science, Springer Verlag.

KEMPEN, G., HOENKAMP, E., (1987), An incremental procedural grammar for sentence formulation, Cognitive science, 11, 201-258.

KENNEDY, C., BOUGAREV, B., (1996), Anaphora for every one: pronominal anaphora resolution without a parser, Proceedings of the 16th International conference on computational linguistics (Coling'96), 113-118, Copenhagen, Denmark.

KERBAT-ORECCHIONI, Catherine, (2001), Oui, Non, Si : un trio célèbre et méconnu, Marges linguistiques, Novembre. http://marges.linguistiques.free.fr/sommaire/sommaire3.htm#b

KERBRAT-ORECCIONI, Catherine, (1996), *La conversation*, Paris : Seuil.

KITA, Kenji, (1996), Mixture Probabilistic Context-Free Grammar : An Improvement of a Probabilistic Context-Free Grammar Using Cluster-Based Language Modeling, *Journal of Natural Language Processing*, Vol.3, No.4, pp.103-113.

KLÜTER, Andreas, *et al.*, (2000), Verbmobil from a software engineering point of view : system design and software integration, *in* Wolfgang Wahlster (éditeur), *Verbmobil : foundations of speech-to-speech translation*, Berlin : Springer.

KRAHMER, E., PIWEK, P., (2000), Varieties of Anaphora, 12th ESSLLI Summer school, Birmingham.

KRENN, Brigitte, SAMUELSSON, Christer, (1997), Statistical Methods in Computational Linguistics, ESSLLI Summerschool, Aix-en-Provence, 1997.

KUIPERS, B. J., (1975), A frame for frames, In D. BOBROW, A. COLLINS (éditeurs), *Representation and understanding :* studies in cognitive science, New York : Academic press.

KURDI, Mohamed-Zakaria, (1996), *Perception phonémique et patrons linguistiques :* étude psycholinguistique du traitement lexical, Mémoire de DES de linguistique de l'université d'Alep, 118 pages.

KURDI, Mohamed-Zakaria, (1999), A Chunk based partial parsing strategy for reranking and normalizing Nbest lists of a speech recognizer, ESSLLI'99, Utrecht, Netherlands.

KURDI, Mohamed-Zakaria, (2000b), La grammaire sémantique d'unification d'arbres : un formalisme pour le l'analyse de la parole spontanée, 7th Conference on Automatic Natural Language Processing TALN'00, Lausanne Swizerland, October 15-18.

KURDI, Mohamed-Zakaria, AHAFHAF, Mohamed, (2002), Toward an objective and generic Method for Spoken Language Understanding Systems Evaluation : an extension of the DCR method, Third International Conference on Language Ressources and Evaluation LREC02, Las Palmas, 29 –31 May.

KURDI, Mohamed-Zakaria., (2000a), Une expertise pour la conception d'un système de classification d'e-mails, Rapport Technique CLIPS-IMAG, Septembre, 101 pages.

LAFFERTY, John, SLEATOR, Daniel, TEMPERLEY, Davy, (1992), Grammatical Trigrams: A Probabilistic Model of Link Grammar, Proceedings of the AAAI Conference on Probabilistic Approaches to Natural Language, October.

LANDAUER, T. K., Foltz, P. W., LAHAM, D., (1998), Introduction to Latent Semantic Analysis, Discourse Processes, 25, 259-284. http://lsa.colorado.edu/

LAPPIN, S., LEAS, H., (1994), An algorithm for pronominal anaphora resolution, Computational linguistics, 20(4), 535-561.

LARREY, P., (1995), *Compréhension du dialogue oral finalisé par une tâche*, rapport de DEA de l'université Paul Sabatier.

LAVIE, Alon, (1997), GLR* : a robust grammar-focused parser for spontaneously spoken language, Ph.D. Dissertation, Carnegie Mellon University, USA.

LEMAIRE, B., (1999), Tutoring Systems based on Latent Semantic Analysis. *9th International Conference on Artificial Intelligence in Education (AIED'99*), Le Mans, 19-23 juillet 1999, published In S.P. Lajoie and M. Vivet (Eds) *Artificial Intelligence in Education, Frontiers In Artificial Intelligence and Applications*, Vol. 50, 1999.

LEVELT, W.J., (1987), *Speaking*: from intention to articulation, Cambrdige, Mass: MIT Press.

LEVELT, W.J.M., (1983), Monitoring and self-repair in speech, Cognition 14, pp41-104.

LEVIN, Lori, GATES, D., LAVIE, A., WAIBEL, A., (1998), An interlingua based domain actions for machine translation of task oriented dialogues, in proceedings of ICSLP-98, *Sydney,* 30th November-4th December.

LICKLEY, Robin L., (1994), *Detecting disfluency in spontaneous speech*, Ph.D. Dissertation, Edinburgh University.

LOPEZ, Patrice, (1999a), *Analyse d'énoncés oraux pour le dialogue homme-machine à l'aide de grammaires lexicalisées d'arbres*, Thèse de L'université Henri Poincaré-Nancy1.

LOPEZ, Patrice, (1999b), Repairing strategies for Lexicalized Tree Grammars, In proceedings of the Ninth Conference of the European Chapter of the Association for Computational Linguistics (EACL'99), Bergen, Norway, June.

LOPEZ, Patrice, Roussel, Davide, (1998), Which rules for robust parsing of spoken utterances with Lexicalized Tree Adjoining Grammars?, Proceedings of the 4rd International Workshop on Tree Adjoining Grammars (TAG+4), Philadelphia, USA. 1-3 August.

LUZZATI, Daniel, (1989), Recherches sur le Dialogue Homme-Machine : modèles linguistiques et traitement automatique, Thèse de doctorat d'état ès lettres, Université de la Sorbonne nouvelle Paris III.

LUZZATTI, Daniel, (1995), *Le dialogue verbal homme-machine* : étude de cas, Paris : Masson.

MACLAY, Howard, OSGOOD, Charles E., (1967), Hesitation phenomena in spontaneous English speech, in L. A. Jakobovits and M. S. Miron (éditeurs), *Readings in the psychology of language*, New Jersey : Prentice-Hall.

MAHESH, Kavi, (1995), *Syntax semantic interaction in sentence understanding*, Ph.D Dissertation, Georgia Institute of technology.

MAYFIELD, Laura, GAVALDÀ, Marsal, WARD, Wayne, WAIBEL, Alex, (1995), Concept based speech translation, Proceedings of the IEEE 1995 International Conference on Acoustics, Speech and Signal Processing (ICASSP 95), Detroit, Michigan, May.

McCLEALAND, James, KAWAMOTO, H., (1986), Machanisms for sentence processing: assigning roles to constituents, in J.L., McCllealand, D., RUMMELHART, (éditeurs), *Parallel and distributed processing*. Exploration in the microstructure of cognition, Vol2: psychological and biological models, Cambridge-Massachussets: the MIT Press.

McCLELLAND, *et al*, (1986), Parallel distributed processing: explorations in the mic rostructure of cognition, Cambridge : MIT press.

McKELVIE, David, (1998), The Syntax of Disfluency in Spontaneous Spoken Language, HCRC Research Paper HCRC/RP-95, May.

METEER, Marie, et al., (1995), Dysfluency annotation stylebook for the Swichboard corpus, Linguistic Data Consortium, http://www.cis.upenn.edu/~treebank/

MILLER, George A., FELLBAUM, Christiane, (1991), Semantic networks of English, In *Cognition* 41, December, pp. 197 - 229.

MILWARD, David, Sturt, Patrick, (1995), Incremental Interpretation, Proceedings of the 7th Conference of the European Chapter of the ACL (EACL), Dublin, Ireland.

MINKER, W., (1995), An English version of the LIMSI L'ATIS system, Rapport technique LIMSI 95-12, Avril.

MINKER, W., BENNACEF, S., GAUVAIN, J.L., (1996), A stochastic case frame approach for natural language understanding, In proceedings of the International Conference on Speech and Language Processing, pages 1013-1016, Philadelphia, October.

MINSKY, M., (1975), A framework for representing knowledge, in Winston, P. (Ed.), *The psychology of computer vision*, New York : McGraw-Hill.

MITKOV, R., (1997), Two engines are better than one : generating more power and confidence in the search for the antecedent, in Ruslan Mitkov and Nicolas Nicolov (editors), *Recent advances in Natural Language Processing*, John Benjamins Publishing Company.

MITKOV, R., (1999), *Anaphora resolution: the state of the art*, Working paper, (Based on the COLING'98/ACL'98 tutorial on anaphora resolution), University of Wolverhampton, Wolverhampton. http ://www.wlv.ac.uk/sles/compling/papers/index.html

MULLER, Claude, (1991), *La negation en français* : Syntaxe, sémantique et éléments de comparaison avec les autres langues romanes, Genève : Droz.

MÜLLER, Stefan, KASPER, Walter, (2000), HPSG analysis of German, *in* Wolfgang Wahlster (ed.), *Verbmobil : foundations of speech-to-speech translation*, Berlin : Springer.

NAKATANI, Christine, HIRSHBERG, Julia, (1994), A Corpus-based study of repair cues in spontaneous speech, JASA 95 :1603-.

NASRI, Mohamed-khaled, (1990), Architecture du système de reconnaissance automatique de la parole DIRA, Thèse de doctorat de l'Institut National Polytechnique de Grenoble, 1990.

NEDERHOF, Mark-Jan, *et al.*, (1997), Grammatical analysis in the OVIS spoken-dialogue system, Proceedings of the ACL/EACL Workshop on Spoken Dialog Systems, Madrid, Spain - July 11-12.

NOY, N., McGUINESS, D. L., (2001), Ontology development 101 : A guide to creating you first ontology, Rapport technique SMI-2001-0880, University of Stanford.

OGDEN, C.K., RICHARDS, I. A., (1923), *The meaning of meaning*, London: Routledge and kegan Paul.

OZKAN, N., (1994), *Analyses communicationnelles de dialogues finalisés*, Thèse Sciences Cognitives, Institut National Polytechnique de Grenoble, Grenoble.

PALOMAR, M., Martinez-Barco, P., (2000), Anaphora resolution through dialogue adjacency pairs and topics, in D-N Christodoulakis, Natural Language Processing - NLP'00, Lecture notes in Artificial Intelligence 1835, Berlin : Springer.

PELLOM, B., WARD, W., HANSEN, J., HACIOGLU, K., ZHANG, J., YU, X., PRADHAN, S., (2001), University of Colorado Dialog Systems for Travel and Navigation, Human Language Technology Conference (HLT-2001), San Diego, March.

PELLOM, B., WARD, W., PRADHAN, S., (2000), The CU Communicator: An Architecture for Dialogue Systems, International Conference on Spoken Language Processing (ICSLP'00), Beijing, China, November.

PEREIRA, Fernando, (1990), Finite-State Approximations of Grammars. In Proceedings of the Second Speech and Natural Language Workshop, pages 20-25, Distributed by Morgan Kaufmann Publishers, San Mateo, California.

PEREIRA, Fernando, WRIGHT, Rebecca N., (1997), Finite-State Approximation of Phrase-Structure Grammars, In Emmanuel ROCHE et Yves SCHABES (éditeurs), *Finite-State Language Processing*, pages 149-173. Cambridge, Massachusetts : MIT Press, 1997.

PERENNOU, Guy, (1996), Compréhension du dialogue oral : Rôle du lexique dans le décodage conceptuel, Séminaire GDR-PRC CHM lexique et communication parlée, octobre.

PICABIA, Lélia, (1975), *Eléments de grammaire générative appliquée au français*, Paris : Armond Colin.

PIERACCINI, R., LEVIN, E., (1991), Stochastic Representation of Semantic Structure for Speech Understanding, In Proceedings of EUROSPEECH 91, Genova, Italy, September.

PIERACCINI, R., LEVIN, E., (1995), A Spontaneous-Speech Understanding System for Database Query Applications, ESCA Workshop on Spoken Dialogue Systems - Theories and Applications, Vigsø, Denmark, May 30, June 2.

PIERACCINI, R., LEVIN, E., LEE, C., (1991), Stochastic representation of conceptual structure in the ATIS task, In Fourth DARPA Workshop on Speech and Natural Language, pages 121--124, Pacific Grove, California, February.

PIERACCINI, Roberto, LEVIN, Esther, ECKERT, Wieland, (1997), AMICA : the ATT Mixed Initiative Conversational Architecture, In Proceedings of Eurospeechù97, pages 1875-1878, Rhodes, Greece, September.

PIERREL, J.M., (1975), Contribution à la reconnaissance automatique du discours continu, Thèse de 3° cycle, Université Nancy I.

PIERREL, J.M., (1978), Myrtille II, un système de compréhension du discours continu, 8° école d'été en Informatique de l'AFCET, Namur, pp. 63-86.

PIERREL, J-M., (1991), Reconnaissance de la parole continue, in Jean Paul Haton, *La reconnaissance automatique de la parole*, Paris : Bordas.

PRIGENT, G., (1994), Synchronous TAGs and machine translation, Proceedinsgs of the third International TAG Workshop, Paris. Rapport technique TALANA-RT-94-01 Université Paris 7.

QIGUANG, L., *et al*, (1997), Key- phrase spotting using an integrated language model of N-grams and finite state grammar, In Proceedings of Eurospeechù97, Rhodes, Greece, September.

QUILLIAN, M.R., (1968), Semantic memory, In M. Minsky, *Semantic information processing*, Cambridge, Massachusetts : MIT Press.

RAPAPORT, William J. (1995), Understanding Understanding: Syntactic Semantics and Computational Cognition, in James E. Tomberlin (ed.), *AI, Connectionism, and Philosophical Psychology*, Philosophical Perspectives Vol. 9 (Atascadero, CA : Ridgeview) : 49-88.

RASTIER, F., (1991), *Sémantique et recherches cognitives*, Paris : P.U.F.

RASTIER, François, (1987), *Sémantique interprétative*, Paris : P.U.F.

RESNIK, P, (1992), probabilistic tree adjoining grammar for statistical natural language processing, In proceedings of COLING'92, Nantes.

ROARK, Brian, JOHNSON, Mark, (1999), Efficient probabilistic top-down and left-corner parsing, In *Proceedings of ACL'99*, pages 421-428.

ROSENBAUM, David, (1987), Neuroscience: the brain and cognition, in N. STILLINGS, Cognitive Science: an introduction, Cambridge-Massachussets: MIT Press.

ROSSI, Mario, *et al.*, (1981), *L'intonation :* de l'acoustique à la sémantique, Paris : Klincksieck, 1981.

ROUILLARD, José, (2000), *Hyperdialogue sur Internet.* Le système HALPIN, thèse de Doctorat de l'université Grenoble I.

ROUSSEL, D., (1999), *Intégration de prédictions linguistiques issues d'applications à partir d'une grammaire d'arbres hors contexte* : Contribution à l'analyse de la parole, Thèse de sciences cognitives, Grenoble, France.

RULAND, Tobias, (2000), Probabilistic LR-parsing with symbolic postprocessing, *in* Wolfgang Wahlster (éditeur), *Verbmobil :* foundations of speech-to-speech translation, Berlin : Springer.

RUPP, C.J., *et al.*, (2000), Combining analyses from various parses, *in* Wolfgang Wahlster (éditeur), *Verbmobil :* foundations of speech-to-speech translation, Berlin : Springer.

SABAH, Gérard, (1989), *L'Intelligence Artificielle et le langage - Tome 2* : processus de compréhension, Paris : Hermès.

SABAH, Gérard, (1990), CARAMEL : un système multi-experts pour le traitement automatique des langues, Modèles linguistiques, 12, Fasc 1, pp. 95-118.

SABAH, Gérard, (1997), Rapport final du projet DALI (1997), http://herakles.imag.fr/pages_html/projets/DALI.html

SABAH, Gérard, RADY, Mohamed, (1983), *A deterministic syntactic-semantic parser applied to French*, Actes 8° IJCAI, Karlsruhe, p. 707-710.

SAHAMI, Mehran, (1998), *Using Machine Learning to Improve Information Access*, Ph.D. Dissertation, Stanford University.

SAIZ-NOEDA, M., PALOMAR, M., (2000), Semantic-driven Knowledge method to solve pronominal anaphora in Spanish texts, in D-N Christodoulakis, Natural Language Processing - NLP'00, Lecture notes in Artificial Intelligence 1835, Berlin : Springer.

SARKAR, Anoop, JOSHI, Aravind, (1996), Coordination in Tree Adjoining Grammars: Formalization and Implementation In the proceedings of COLING 1996, Copenhagen. pp. 610-615.

SAUVAGEOT, Aurélien, (1972), *Analyse du français parlé*, Paris : Hachette.

SCHABES, Y., (1992), Stochastic lexicalized tree adjoining grammars, In proceedings of COLING'92, Nantes.

SCHABES, Y., WATERS R.C., (1995), Tree Insertion Grammar: A Cubic-Time Parsable Formalism That Lexicalizes Context-Free Grammar Without Changing the Trees Produced, Computational Linguistics, 21(4) :479--513, December 1995.

SCHANK, R. C., (1977), Conceptual dependency: a theory of natural language understanding, Cognitive psychology, 3, 4, pp. 552-630.

SCHANK, R. C., ABELSON, R., (1977), Scripts, plans, goals and understanding, Hillsdale (N.J.): Erlbaum.

SCHEEPERS, C., CORLEY, Martin, (2000), Syntactic priming in German sentence production, in the proceedings of the Twenty Second Annual Meeting of the Cognitive Science Society CogSci2000,, Philadelphia, Pennsylvania, USA, August 13-15.

SCHIEHLEN, Michael, *et la.*, (2000), Verbmobil Interface Terms (VITs), *in* Wolfgang Wahlster (éditeur), *Verbmobil :* foundations of speech-to-speech translation, Berlin : Springer.

SCHWARTZ, J. L., (1996), Perception de la parole : des représentations sensorimotrices à l'émergence des systèmes linguistiques, in MELONI, H., (éditeur principal), *Fondements et perspectives en traitement automatique de la parole*, Paris : Aupelf-UREF.

SEARLE, J. R., (1996), *Les actes de langage, essai de philosophie du langage*, Paris : Hermann, 1996.

SÉRASSET, Gilles, BOITET, Christian, (1999), UNL-French deconversion as transfer & generation from an interlingua with quality enhancement though offline human interaction, MT Summit 99, 13-17 September, Singapore, pp 220-228.

SHIEBER, Stuart M., SCHABES, Yves, PEREIRA, Fernando C. N., (1995), Principles and implementation of deductive parsing. *Journal of Logic Programming*, volume 24, number 1-2, pages 3-36.

SHIEBER, Stuart, Schabes, Yves, (1990), Synchronous Tree-Adjoining Grammars. In Proceedings of the 13th International Conference on Computational Linguistics, volume 3, pages 1-6.

SHRIBERG, E., (1994), *Preliminaries to a theory of speech disfluencies*, Ph.D. Dissertation, University of Berkeley.

SIEGEL, Melanie, (2000), HPSG analysis of English, *in* Wolfgang Wahlster (éditeur), *Verbmobil:* foundations of speech-to-speech translation, Berlin : Springer.

SIKKEL, Klaas, (1997), *Parsing schemata*, Berlin : Springer-Verlag, 1997.

SOWA, J., (1982), Using a lexicon of canonical graphs in a semantic interpreter, in WALTON EVENS, M., *Relational model of the lexicon : representing knowledge in semantic networks*, Cambridge: Cambridge university press.

SPERBER, Dan, WILSON, Deirdre, (1995), *Relevance:* Communication and Cognition, Oxford : Blackwell.

SPIVEY-KNOWLTON, M., TANENHAUS, M., Immediate effect of discourse and semantic context in syntactic processing: Evidence from eye tracking, in proceedings of the fifteenth annual conference of cognitive science society pages 812 – 817.

SRINIVAS, B., (1996), Almost Parsing Technique for Language Modeling, International Conference on Speech and Language Processing (ICSLP-96), Philadelphia, PA, USA, October, 1996.

SRINIVAS, B., (1997), *Complexity of lexical description and its relevance to partial parsing*, Ph.D. Dissertation, University of Pennsylvania.

STAAB, Steffen, (1994), *GLR-Parsing of Word Lattices*, Master's thesis, University of Pennsylvania.

TRUSWELL, J.C., TANENHAUS, M. K., (1992), Consulting temporal context during sentence comprehension: evidence from the monitoring of eye movements in reading, in proceedings of the fourteenth annual conference of cognitive science society, pages 492 – 497.

TYLER, L. K., MARSLEN-WILSON, W. D. (1977), The on-line effects of semantic context on syntactic Processing, Journal of Verbal Learning and Verbal Behavior, 16:683-692.

UCHIDA, H., ZHU, M., (2001), The Universal Networking Language beyond machine translation, International Symposium on Language in Cyberspace, 26 - 27 September 2001, Seoul, Korea http://www.unl.ias.unu.edu/publications/index.html.

ULRICH, Hans, RULAND, Tobias, (2000), Integrated shallow processing, *in* Wolfgang Wahlster (ed.), *Verbmobil : foundations of speech-to-speech translation*, Berlin : Springer.

USZKOREIT, H., *et al.*, (2000), Deep linguistic analysis with HPSG, In Wolfgang Wahlster (éditeur), *Verbmobil :* foundations of speech-to-speech translation, Berlin : Springer.

VAN NOORD, Gertjan, (1997), An Efficient Implementation of the Head-Corner Parser, Computational Linguistics, volume 23, number 3. http://odur.let.rug.nl/~vannoord/papers/

VAUFREYDAZ, Dominique, BERGAMINI, Carole, SERIGNAT, Jean-François, AKBAR, Mohamad, (2000), A New Methodology for Speech Corpora Definition from Internet Documents, In proceedings of the International Conference on Language Ressources and Evaluation LREC`2000, 2nd International Conference on Language Resources and Evaluation, Athens, Greece, 3 vol. p.423-426.

VILLANEAU, Jeanne, ANTOINE, Jean-Yves, RIDOUX, Olivier, (2002), LOGUS : un système formel de compréhension du français parlé spontané - présentation et évaluation, TALN'2002. Nancy, France. Juin 2002.

WAHLSTER, Wolfgang (éditeur), (2000), *Verbmobil :* foundations of speech-to-speech translation, Berlin : Springer.

WAHLSTER, Wolfgang, (2000), Mobile speech-to-speech translation of spontaneous dialogs: an overview of the final Verbmobil system, in Wolfgang Wahlster (éditeur), *Verbmobil: foundations of speech-to-speech translation*, Springer.

WANG, Ye-Yi, (2001), Robust language understanding in Mipad, In proceedings of Eurospeech'01, Aalborg, Denmark, 3-7 September.

WARD, Wayne, (1991), Understanding spontaneous speech: the phoenix system, In proceedings of International conference on Acoustics, Speech and signal processing, pages 365-367, May.

WIEMER-HASTINGS, P., K., et al., (1999), Improving an intelligent tutor's comprehension of students with Latent Semantic Analysis, In S. Lajoie and M. Vivet, (éditeurs), Artificial Intelligence in Education, Amsterdam, pages 535-542.

WILLIAMS, Edwin, (1994), *Thematic roles in Syntax*, Cambridge, Mass.: MIT Press.

WOODS, Z. A., (1970), Transition network grammar for natural language analysis, CACM 13 (10): 591-606, Reproduit dans B. J. Grosz et al. (éditeurs), *Readings in Natural Language Processing*, Los Altos: Morgan Kaufmann Publishers, 1986.

WOSZCZYNA, M., BROADHEAD, M., GATES, D., GAVALDÀ, M., LAVIE, A., LEVIN, L., WAIBEL, A., (1998), A Modular Approach to Spoken Language Translation for Large Domains, Proceedings of the AMTA.'98.

ZEC, D., INKELAS, S., (1995), *Phonology syntax connection*, Chicago: Chicago University Press.

ZECHNER, C., WAIBEL, A., (1998), Using chunk based partial parsing of spontaneous speech in unrestricted domains for reducing word error rate in speech recognition, COLING-ACL'98, August 10-14, Montréal.

2 Bibliographie générale[60]

ABEILLE, Anne, GODARD, Danièle, MILLER, Philip, SAG, Ivan A., (1998), French Bounded Dependencies, In Luca Dini and Sergio Balari, (éditeurs), Romance in HPSG. Stanford: CSLI Publications.

ABEILLÉ, Anne, GODARD, Danièle, SAG., Ivan A., (1998), Two Kinds of Composition in French Complex Predicates, In Erhard Hinrichs, Andreas Kathol, Tsuneko Nakazawa, (éditeurs), Complex Predicates in Nonderivational Syntax, New York: Academic Press.

ABNEY, Steven, (1990), Rapid Incremental Parsing with Repair, In: Proceedings of the 6th New OED Conference: Electronic Text Research, pp.1-9. University of Waterloo, Waterloo, Ontario, October.

ABNEY, Steven, (1992), Prosodic Structure, Performance Structure and Phrase Structure, In: Proceedings, Speech and Natural Language Workshop, pp.425-428. Morgan Kaufmann Publishers, San Mateo, CA. 4 pages, 37 KB.

ABNEY, Steven, (1994), Partial Parsing, Tutorial given at ANLP-94, Stuttgart, October 1994. http://www.sfs.nphil.uni-tuebingen.de/~abney/Papers.html#96i.

ABNEY, Steven, (1995), Chunks and Dependencies: Bringing Processing Evidence to Bear on Syntax. In: Computational Linguistics and the Foundations of Linguistic Theory. CSLI.

ABNEY, Steven, (1996), Statistical Methods and Linguistics, In Judith Klavans, Philip Resnik (éditeurs), *The Balancing Act*, Cambridge, MA : The MIT Press.

ABNEY, Steven, (1996), Tagging and Partial Parsing, In: Ken Church, Steve Young, and Gerrit Bloothooft (eds.), Corpus-Based Methods in Language and Speech, An ELSNET volume. Kluwer Academic Publishers, Dordrecht. 23+1 pages.

AHO, Alfred V., SETHI, Ravi, ULLMAN, Jeffrey, (1988), *Compilers*: principles, techniques and tools, Reading, Massachusetts : Addison-Wesley.

ALSHAWI, *et al.*, (1988), Research Programme in Natural Language Processing : July 1988 Annual Report, SRI International Tech Note, Cambridge England.

ANTOINE, Jean Yves, *et al.*, (1998), DQR Test suites for a qualitative evaluation of spoken dialog systems: from speech understanding to dialog strategy, International Conference on Language Ressources and Evaluation LREC'98, Grenade, Espagne.

ARAGUES PELEATO, R., RAJMAN, M., CHAPPELIER, J.-C., (1999), Integration of syntactic constraints within a speech recognition system: Coupling a speech recognizer and a stochastic context-free parser, Technical Report No 98/309, Département Informatique, EPFL, Lausanne, Switzerland, February 23.

BAMHAMED, M., (1995), *Traitement en temps réel des énoncés complexes :* étude comparative intralangues, Thèse de doctorat en psychologie expérimentale présentée à l'université Paris V.

BARRIERE, C., (1997), From children's first dictionary to a lexical knowledge base of conceptual graphs, Ph.D. Simon Fraser university.

BECHET, F., *et al*, (1994), Détection de mots Clés dans un discours continu, XXèmes Journées Francophones d'Etudes sur la Parole JEP.

[60] Cette liste contient les références que nous avons explorées avant ou au cours de la préparation de cette thèse.

BENSON TESAR, B., (1995), *Computational optimal Theory*, Ph.D. Dissertation, University of Colorado.

BIGI, B., DE MORI, R., EL-BÈZE, M., SPRIET, T., (1997), Combined models for topicspotting and topic-dependent language modeling, ASRU97, Santa Barbara, USA, décembre.

BLACK, Ezra, JELINEK, Fred, LAFFERTY, John, MAGERMAN, David M., MERCER, Robert, ROUKOS, Salim, (1992), Towards History-based Grammars: Using Richer Models for Probabilistic Parsing, in Proceedings, DARPA Speech and Natural Language Workshop, February.

BOD, R., (1996), Efficient Algorithms for Parsing the DOP Model? A Reply to Joshua Goodman, Computational Linguistics Archive: cmp-lg/9605031.

BOD, R., (1996), Two Questions about Data-Oriented Parsing, Proceedings Fourth Workshop on Very Large Corpora, COLING'96, Copenhagen, Denmark.

BOD, R., BONNEMA, R., SCHA, R., (1996), A Data-Oriented Approach to Semantic Interpretation. Proceedings Workshop on Corpus-Oriented Semantic Analysis, ECAI-96, Budapest, Hungary.

BOD, R., SCHA, R., (1997), Data-Oriented Language Processing, In S. Young and G. Bloothooft (éditeurs), Corpus-Based Methods in Language and Speech Processing, Kluwer Academic Publishers, Boston. 137-173.

BODIN, N., *et al*, (1997), Analyse lexicale en vue de la génération de la parole, Maîtrise MASS de l'Université Grenoble II.

BOITET, Christian, (1998), Problèmes scientifiques intéressants en traduction de parole, colloque NLP+IA-98, Moncton, 18-21 août.

BORILLO, A., (1993), Préposition de lieu et anaphore, Langages110, juin 1993.

BOROS, M., ARETOULAKI, M., GALLWITZ, F., NIEMANN, H., NÖTH, E., (1997), Semantic Processing of Out-of-Vocabulary Words in a Spoken Dialogue System. In *Proceedings of the 5th European Conference on Speech Communication and Technology (EUROSPEECH'97)*, volume 4, pages 1887-1890, Rhodes, Greece.

BOROS, Manuela, ARETOULAKI, Maria, GALLWITZ, Florian, NÖTH, Elmar, NIEMANN, Heinrich, (1997), Semantic Processing of Out-of-Vocabulary Words in a Spoken Dialogue System, In: Proceedings of EUROSPEECH'97, pp. 1887-1890, Rhodes, Greece.

BOUAUD, J., *et al*, (1992), A reconstruction of conceptual graphs on top of a prodution system, Las Cruces (NM), 7 Th. annual workshop on conceptual structures, juillet.

BOUAUD, J., *et al*, (1996), Processing metonymy a domain -model heuristic graph traversal approach, Copenhagen: COLING, Août.

BOUMA, Gosse, KOELING, Rob, NEDERHOF, Mark-Jan, VAN NOORD, Gertjan, (1996), Conventional Natural Language Processing in the NWO Priority Programme on Language and Speech Technology, October 1996 Deliverables., 13 nov-96, Groningen.

BOUMA, Gosse, Rob, Koeling, Mark-Jan, Nederhof, Gertjan, van Noord, (1996), Conventional Natural Language Processing in the Priority Programme on Language and Speech Technology: January 1996 Deliverables, Groningen.

BOUMA, Gosse, Rob, Koeling, Mark-Jan, Nederhof, Gertjan, van Noord, (1996), Grammatical Analysis in a Spoken Dialogue System, Groningen. http://odur.let.rug. nl:4321/publijst.html.

BRATT, H., DOWDING, D., HUNICKE-SMITH, K, (1995), The SRI Telephone-based ATIS System, Proceedings of the Spoken Language Systems Technology Workshop, Austin, Texas, January.

BYRON, Donna K., ALLEN, James F., (1998), Resolving Demonstrative Anaphora in the TRAINS93 Corpus, In Proceedings of DAARRC2 - Discourse, Anaphora and Reference Resolution Colloquium, Lancaster University, August.

BYRON, Donna K., HEEMAN, Peter A., (1997), Discourse Marker Use in Spoken Dialog, In Proceedings of the 5th European Conference On Speech Communication and Technology, Rhodes, Greece, September, pages 2223-2226.

CAELEN, J., (1996), Architecture logicielle en reconnaissance, parallélisme et modularité, in MELONI, H., (éditeur principal), Fondements et perspectives en traitement automatique de la parole, Paris: Aupelf-UREF.

CALLIOPE, (1989), *La parole et son traitement automatique*, Paris: Masson.

CANDITO, M. H., (1999), Orgnisation modulaire et paramétrable de grammaires éléctroniques lexicalisées : Application au français et à l'italien, Thèse de doctorat de l'université Paris 7.

CARON, Jean, (1983), *Les régulations du discours :* psycholinguistique et pragmatique du langage, Paris : Presses Universitaires de France.

CARROLL, J., *et al*, (1997), SPARKLE: Work Package 1 specification of phrasal parsing, Final report, November.

CARTER, David, RAYNER, Manny, (1994), The speech-language interface in the Spoken Language Translator", Proceedings of TWLT-8, Twente Workshop on Language Technology, University of Twente, Holland, December 1994.

CETTOLO, M., CORAZZA, A., De MORI, R., (1996), A Mixed Approach to Speech Understanding, Proceedings of ICSLP 96, International Conference on Spoken Language Processing, Philadelphia, PA, USA, 3rd-6th October.

CHANOD, Jean-Pierre, Tapanainen, Pasi, (1994), Tagging French - Comparing a Statistical and a Constraint-based Method, (Nov 1994) MLTT-016.

CHARAUDEAU, Patrick, (1992), *Grammaire du sens et de l'expression*, Paris: Hachette.

CHARNIAK, Eugene, (1997), Statistical parsing with a context-free grammar and word statistics, Proceedings of the Fourteenth National Conference on Artificial Intelligence AAAI Press/MIT Press, Menlo Park.

CHARNIAK, Eugene, (1997), Statistical techniques for natural language parsing AI Magazine.

CHEN, Stanley, (1996), *Building probabilistic model for natural language*, Ph.D. Dissertation, Harvard University.

CHICOISNE, Guillaume, (1998), *Conversation et relations sociales pour des agents moins artificiels*, DEA Sciences Cognitives.

CIRAVEGNA, Fabio, LAVELLI, Alberto, (1995), On Parsing Control for Efficient Text Analysis, Proceedings of the Fourth International Workshop on Parsing Technologies, Prague, Czech Republic, September.

CIRAVEGNA, Fabio, LAVELLI, Alberto, (1997), Controlling Bottom-Up Chart Parsers through Text Chunking in Proceedings of the 5th International Workshop on Parsing Technologies (IWPT97), Boston, September, 17-20.

CLARK, Stephen, WEIR, David, (1999), An Iterative Approach to Estimating Frequencies over a Semantic Hierarchy, in *Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora.*

COLINEAU N., A. HALBER, (1999), Hybrid approach to spoken query processing in document retrieval system", ESCA ETRW Workshop: Accessing Information in Spoken Audio, Cambridge, UK.

COLINEAU, Nathalie, *et al*, (1996), Une approche lexicale pour la reconnaissance d'actes de dialogue, Séminaire GDR-PRC CHM lexique et communication parlée, octobre.

COLINEAU, Nathalie, HALBER, Ariane, (1999), Une approche hybride pour l'interrogation en langage naturel parlé de bases de données documentaires, l'Atelier Intégration Parole et Langage,TALN'99, Corse.

CONIAM, D., (1998), Partial parsing: Boundary marking International journal of corpus linguistics, vol. 3, no 2, pp. 229 – 249, Benjamins publishing, Amsterdam.

CORAZZA Anna, LAVELLI, Alberto, (1994), An N-Best Representation for Bidirectional Parsing Strategies, In Working Notes of the AAAI'94 Workshop on the Integration of Natural Language and Speech Processing, Seattle, WA, August.

CORAZZA, Anna, *et al*, (1999), The ITC-IRST Speech translation system, C-Star Workshop, Schwetzingen, Germany, September 23-24, 1999.

CORE, Mark, SCHUBERT, Lenhart, (1999), A Syntactic Framework for Speech Repairs and Other Disruptions, Presented at the 37th Annual Meeting of the Association for Computational Linguistics (ACL-99), College Park, MD, June.

CORE, Mark, SCHUBERT, Lenhart, (1999), Speech Repairs: A Parsing Perspective, Presented at the ICPhS Satellite Meeting on Disfluency in Spontaneous Speech, Berkeley, CA, July-99.

CORE, Mark, SCHUBERT, Lenhart, (1999), A Model of Speech Repairs and Other Disruptions, Presented at AAAI Fall Symposium on Psychological Models of Communication in Collaborative Systems, Cape Cod, MA, November.

CORI, Marcel, et al., (1997), Parsing repair, in Ruslan Mitkov and Nicolas Nicolov (editors), *Recent advances in Natural Language Processing*, John Benjamins Publishing Company.

COVINGTON, Michael A., (1990), A dependency parser for variable word order languages, AI-1990-01, The university of Georgia, Athens Georgia.

COVINGTON, Michael A., (1990), An empirically motivated reinterpretation of dependency grammar, Research report AI-1994-01, The university of Georgia, Athens Georgia.

COVINGTON, Michael A., (1994), Discontinues dependency parsing of free and fixed word order, Research report AI-1994-02, The university of Georgia, Athens Georgia.

CREISSEILS, Denis, (1995), *Eléments de syntaxe générale*, Paris : PUF.

CRISTEA, Dan, (2000), An incremental discourse parser architecture, in D-N Christodoulakis, Natural Language Processing - NLP'00, Lecture notes in Artificial Intelligence 1835, Berlin: Springer.

CRISTEA, Dan, Bonnie, WEBBER, (1997), Expectations in Incremental Discourse Processing. Proc. 35th Annual Meeting of the Association for Comptutational Linguistics, Madrid, July 1997.

CROUCH, R., (1995), Ellipsis and Quantification: a substitutional approach, Proceedings of the 7th European ACL, Dublin, Ireland, February Report CRC-054

DAHAN, D., (1994), Etude de l'influence des variations prosodiques dans les processus de traitement de la parole, Trégastel, XX èmes JEP, Juin.

DEJEAN, Hervé, (1998), *Concepts et algorithmes pour la découverte des structures formelles des langues*, thèse de doctorat, Université de Caen.

DEROUARD, Laurent, (1997), Traitement de la parole spontanée par réseaux récurrents, Mémoire de DEA Sc. cognitives, INPG.

DOWDING, Gawron, *et al.*, (1993), GEMINI: A Natural Language System for Spoken-Language Understanding, Proc. of the 31st Annual Meeting of the Association for Computational Linguistics.

DOWDING, J., GAWRON, M., APPELT, J., BEAR, D., CHERNY, J., MOORE, L., MORAN, D., (1993), GEMINI: A natural language system for spoken-language understanding, in Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics, (Ohio State University, Columbus, Ohio), pp. 54--61, 22--26 June.

DRESHER, B. E., (1996), Introduction to metrical and prosodic Phonology, in MORGAN, J., *et al*, (sous la direction), Signal to Syntax: bootstrapping from speech to grammar in early acquisition, Mahwah: Laurence Erlbaum associates.

DUBOIS, D., (1991), Catégorisation et cognition <<18 ans après>>: une évaluation des concepts de Rosch, in DUBOIS, D., Sémantique et cognition: catégories et prototypes, typicalité, Paris: Editions du CNRS.

DUPONT, Pierre, (1996), *Utilisation et apprentissage de modèles de langage pour la reconnaissance de la parole continue*, Thèse de doctorat, ENST Paris.

ECHOLS, C. H., (1996), A Role for stress in early speech segmentation, in MORGAN, J., *et al*, (sous la direction), Signal to Syntax: bootstrapping from speech to grammar in early acquisition, Mahwah: Laurence Erlbaum associates.

ECKERT, M., (1998), *Discourse Deixis and Null Anaphora in German*, Ph.D. Dissertation, Edinburgh University.

ECKERT, M., STRUBE, M., (1999), Dialogue Acts, Synchronising Units and Anaphora Resolution, In: Proceedings of Amstelogue'99, Workshop on the Semantics and Pragmatics of Dialogue, Amsterdam University, 7-9 May.

ECKERT, Wieland, NIEMANN, Heinrich, (1994), Semantic Analysis in a Robust Spoken Dialog System, In Proc. International Conference on Spoken Language Processing, pages 107-110, Yokohama.

EHRLICH, M. F., *et al*, (1993), *Les modèles mentaux: approche cognitive des représentations des connaissances*, Paris: Masson.

EL-BEZE, M., SPRIET, T., (1995), Intégration de contraintes syntaxiques dans un Système d'étiquetage probabiliste, revue TAL, décembre.

FARGUES, Jean, *et al*, (1986), Conceptual graphs for semantic and knwoledge processing, IBM J. RES. DEVELOP VOL 30 No.1 January.

FOSTER, Jennifer, (2000), Feature structures and syntactic inconsistency, ESSLLI'00 Student session, Birmingham, UK.

FRITSCH, J., (1996), *Modular neural network for speech recognition*, Ph.D Dissertation, CMU.

GABI, K., (1997), Extraction dynamique de connaissances à partir de textes par réseaux neuronaux, Mémoire de DEA, INP Grenoble.

GARDE, P., (1968), *L'Accent*, Paris: Presses Universitaires de France PUF.

GAUVAIN, J.L., BENNACEF, S.K., DEVILLERS, L., LAMEL, L.F., and ROSSET, S., (1997), Spoken language component of the MASK kiosk. In K. Varghese and S. Pfleger, editors, *Human Comfort and Security of Information Systems*, pages 93-103. Springer.

GAVALDÀ, Marsal, (2000), Epiphenomenal Grammar Acquisition with GSG, In Proceedings of the Workshop on Conversational Systems of the 6th Conference on Applied Natural Language Processing and the 1st Conference of the North American Chapter of the Association for Computational Linguistics (ANLP/NAACL-2000), Seattle, Washington, U.S.A., May 2000.

GAVALDÀ, Marsal, (2000), SOUP: A Parser for Real-world Spontaneous Speech, In Proceedings of the Sixth International Workshop on Parsing Technologies (IWPT-2000), Trento, Italy, February 2000.

GERARD, C., (1994), Patrons prosodiques et intentions des locuteurs: production et perception expressive chez l'adulte et chez l'enfant, Trégastel, XX èmes JEP, Juin.

GHIGLIONE, R., TROGNON, A., (1993), Où va la pragmatique, Presses universitaires de Grenoble.

GHORBEL, H., PALLOTTA, V., (2000), Weighted Robust Parsing Approach to Semantic Annotation, In proceedings of the ANLP-NAACL'2000, Student Workshop, p 19-23, April 29-May 4, Seattle, Washington, USA.

GIGUET, Emmanuel, (1998), *Méthode pour l'analyse automatique de structures formelles sur documents multilingues*, thèse de doctorat de l'université de Caen, 1998.

GIGUET, Emmanuel, VERGNE, Jacques, (1997), From Part-of-Speech Tagging to Memory-based Deep Syntactic Analysis, In Proceedings of the International Workshop on Parsing Technologies (IWPT'97), MIT, Boston, Massachussets, USA, September 17-20.

GIGUET, Emmanuel, VERGNE, Jacques, (1997), Syntactic analysis of unrestricted French. In proceedings of the International Conference on Recent Advances in Natural Languages Processing (RANLP'97), pages 276-281, Tzigov Chark, Bulgaria, September 11-13.

GOLDING, Andrew R., SCHABES, Yves, (1996), Combining Trigram-based and Feature-based Methods for Context-Sensitive Spelling Correction, In Proceedings of the 34th Annual Meeting of the Association for Computational Linguistics, Santa Cruz, CA.

GOODMAN, J., (1996), Efficient algorithms for parsing the DOP model. In Proceedings of the Conference on Empirical Methods in Natural Language Processing, pages 143-152, May.

GRAESSER, A. C., *et al*, (sans date), Using latent semantic analysis to evaluate the contribution of students in autotutor, Submition for Special Issue of Interacting Learning Environments, guest edited by Joseph Psotka. http://Mnemosyne.csl.psyc.memphis.edu/home/graesser/PSOTKA.htm.

GRUBER, T. R., (1993), Toward principles for the design of ontologies used for knowledge sharing, International Workshop on Formal Ontology, March 1993. Available as Stanford Knowledge Systems Laboratory Report KSL-93-04.

HAAS, J., HORNEGGER, J., NIEMANN, H., (1998), Probabilistic Semantic Analysis In Restricted Domains. In Proceedings of the International Workshop on Speech and Computer, pages 151-158, St. Petersburg, Rußland, October.

HAAS, J., WARNKE, V., NIEMANN, H., CETTOLO, M., CORAZZA, A., FALAVIGNA,

HACIOGLU, Kadri, WARD, Wayne, (2001), A Word Graph Interface for a Flexible Concept Based Speech Understanding Framework, Eurospeech'2001, Aalborg Denmark, September.

HALBER A., (1998), Grammatical Factor and Spoken Sentence Recognition" TSD'98, Workshop on Text, Speech and Dialogue article.

HARPER M. P., et al., (1998), Interfacing Acoustic Models with Natural Language Processing Systems Proceedings of the International Conference on Spoken Language Recognition ICSLP'98, Sydney, Australia.

HATON, J. P., et al, (1991), *Reconnaissance automatique de la parole*, Paris : DUNOD.

HEEMAN, Peter A., ALLEN, James F., (1994), Tagging Speech Repairs, In ARPA Workshop on Human Language Technology, pages 187-192, Princeton, March,

HEEMAN, Peter, and Kyung-ho Lokem-Kim, (1995), Using structural information to detect speech repairs, TR IEICE SP95-91, December.

HEYD, S., (1995), *La détermination nomilale en français* : étude dans le formalisme HPSG de deux phénomènes particuliers, mémoire de DEA de l'université Nancy II.

HOFFMANN, Marc, LANG, Manfred, (2000), Belief networks for a syntactic and semantic analysis of spoken utterances for speech understanding, In the proceedings of the 6th international conference on spoken language processing (ICSLP'00), Beijing, China.

HOFSTADTER, Douglas R., (1979), *Goedel, Escher, Bach: an Eternal Golden Braid*, NY: Basic Books.

BRANIGAN, H. P., (1995), Language Processing and the Mental Representation of Syntactic Structure (1995, 252 pages), EUCCS-PHD-1995-8.

http://www.smi.stanford.edu/projects/protege/publications/ontology_development/ontology101.html

HUDSON, Richard, (1998), Grammar without functional categories, in R. Borsley (éditeur) *The Nature and Function of Syntactic Categories*, London, Academic Press.

HURST, Mathew F., (1997), Parsing for targeted errors in controlled languages, in Ruslan Mitkov and Nicolas Nicolov (editors), *Recent advances in Natural Language Processing*, John Benjamins Publishing Company.

IIDA, Hitochi, (1999), Speech translation for primitive conversations, C-Star Workshop, Schwetzingen, Germany, September 23-24.

ISSAC, Fabrice, (1997), Analyse syntaxique et apprentissage des langues, Thèse de doctorat de l'université Paris-Nord.

ISSAR, Sunil, WARD, Wayne, (1994), Integrating Semantic Constraints Into The SPHINX-II Recognition, Volume: 2 Page 17 Paper number 2017.

ISSAR, Sunil, WARD, Wayne, (1994), Unanswerable Queries In A Spontaneous Speech Task, ICASSP, Volume: 1 Page 341 Paper number 1341

JOKINEN, Kristiina, (1995), Communicative Principles and Utterance Planning, Proceedings of The 4th International Conference on The Cognitive Science of Natural Language Processing, July-95.

JOSHI, Aravind, Some linguistic, computational and statistical implications of Lexicalized grammars, in Ruslan Mitkov and Nicolas Nicolov, Recent advances in Natural Language Processing, John Benjamins Publishing Company, 1997.

JUNQUA, Jean-clause, HATON, Jean-Paul, (1996), *Robustness in automatic speech recognition : fundamentals and applications*, Kluwer academic publishers, Boston.

JURAFSKY, D., BATES, R., COCCARO, N., MARTIN, R., METEER, M., RIES, K., SHRIBERG, E., STOLCKE, A., TAYLOR, P., Ess-Dykema C. Van, (1997), Automatic Detection of Discourse Structure for Speech Recognition and Understanding. Proc. IEEE Workshop on Speech Recognition and Understanding, 88-95, Santa Barbara, CA.

JURAFSKY, D., C. Wooters, J. Segal, A. Stolcke, E. Fosler, G. Tajchman, & N. Morgan (1995), Using a Stochastic Context-Free Grammar as a Language Model for Speech Recognition. Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing,189-192, Detroit.

KAY, Martin, (1980), Algorithm schemata and data strucutres in syntactic processing, Readings in natural language processing – Moragan Kaufman publishers, Los Altos, California, 35 – 70.

KEHLER Andrew, SHIEBER, Stuart, Anaphoric Dependencies in Ellipsis. Computational Linguistics, volume 23, number 3, 1997.

KEHLER Andrew, (1993), A Discourse Copying Algorithm for Ellipsis and Anaphora Resolution, In Proceedings of the Sixth Conference of the European Chapter of the Association for Computational Linguistics (EACL-93), pp. 203-212, Utrecht, April.

KELLER, Frank, (1996), How Do Humans Deal with Ungrammatical Input? Experimental Evidence and Computational Modelling, In Dafydd Gibbon, ed., Natural Language Processing and Speech Technology: Results of the 3rd KONVENS Conference, Bielefeld, October 1996, 27--34. Berlin: Mouton de Gruyter.

KELLER, Frank, (1998), Gradient Grammaticality as an Effect of Selective Constraint Re-Ranking. To appear in M. Catherine Gruber, Derrick Higgins, Kenneth Olson, and Tamara Wysocki, eds., Papers from the 34th Annual Meeting of the Chicago Linguistic Society. Vol. 2: The Panels. Chicago.

KELLER, Frank. (1997), Extraction, Gradedness, and Optimality, In Alexis Dimitriadis, Laura Siegel, Clarissa Surek-Clark, and Alexander Williams, (éditeurs), Proceedings of the 21st Annual Penn Linguistics Colloquium, 169-186. Penn Working Papers in Linguistics 4.2, Department of Linguistics, University of Pennsylvania.

KEMPSON, R., MEYER VIOL, W., and GABBAY, D. (forthcoming), Syntactic Computation as Labelled Deduction: WH a case study, in R., BORSLEY, I., ROBERTS (éditeurs), *Syntactic Categories*, Academic Press.

KIM, A., SRINIVAS, B.J., TRUESWELL, J., (1997), Incremental Processing Using Lexicalized Tree-Adjoining Grammar: Symbolic and Connectionist Approaches, Conference on Computational Psycholinguistics, Berkeley, California, August.

KING, Margaret, (1995), The Evaluation of NLP Systems, In: Swan21 Proceedings, Geneva, pp. 97-109.

KINYON, A., (1998), Un algorithme d'analyse LR(0) pour les grammaires d'arbres Adjoints Lexicalisées, TALN'98, Paris, 10-12 Juin.

KITA, Kenji, (1996), Mixture probabilistic context free grammar: an improvement of a probabilistic context free grammar using cluster based language modeling, Journal of natural language processing Vol. 3, No. 4, October.

KITA, Kenji, MORIMOTO, Tsuyoshi, OHKURA, Kazumi, SAGAYAMA, Shigeki, (1992), Continuously Spoken Sentence Recognition by HMM-LR, 2nd International Conference on Spoken Language Processing, pp.305-308.

KLEIBER, G., (1990), *La Sémantique du prototype*: catégories et sens lexical, Paris : PUF.

KLEIBER, G., (1991), Prototype et prototype: encore une affaire de famille, *in* DUBOIS, D., *Sémantique et cognition* : catégories et prototypes, typicalité, Paris: Editions du CNRS.

KNOTT, Alistair, (1996), *A data-driven methodology for motivating a set of coherence relations*, Ph.D. Dissertation, University of Edinburgh.

KNUTH, Donald E., (1997), *The art of computer programming*, Volume 1Fundamental algorithms, Reading: Addison-Wesley.

KODRATOFF, Y., (1997), L'extraction des connaissances à partir des données: un nouveau sujet pour la recherche scientifique, READ volume 1 (1).

KORBAYOVÁ, Ivana, KRUIJFF, Geert-Jan, (1996), Identification of topic-focus chains, In S., Botley, *et al.*, (éditeurs), Approaches to Discourse Anaphora: Proceedings of the Discourse Anaphora and Anaphora Resolution Colloquium (DAARC96), number 8 in Technical papers, pages 165--179, Lancaster, U.K.

KRAHMER, E., VAN DEEMTER, K., (1998), On the Interpretation of Anaphoric Noun Phrases: Towards a Full Understanding of Partial Matches. In: Journal of Semantics, 15(3/4), 355-392.

KRAHMER, E., (1996), Presuppositional Discourse Representation Theory, In Proceedings of the Tenth Amsterdam Colloquium, P. Dekker and M. Stokhof (eds.), ILLC, Amsterdam, 499-518.

KRAHMER, E., VAN DEEMTER, K., (1997), Presuppositions as Anaphors: Towards a Full Understanding of Partial Matches. In: De Dag, Proceedings of the Workshop on Definites, P. Dekker, J. van der Does, H. de Hoop (eds.), Utrecht Institute of Linguistics, 81-112.

KRAHMER, E., (1995), *Discourse and Presupposition*, Ph.D Dissertation, Tilburg University.

KRAHMER, E., K., van, Deemter, (1997), Partial Matches and the Interpretation of Anaphoric Noun Phrases. In: Proceedings of the 11th Amsterdam Colloquium, P. Dekker and M. Stokhof (eds.), ILLC, Amsterdam, 205-210.

KRAHMER, E., LANDSBERGEN, J., POUTEAU, X., (1997), How to Obey the 7 Commandments for Spoken Dialogue Systems. In: Proceedings of the (E)ACL workshop on Interactive Spoken Dialog Systems, J., HISCHBERG, C. KAMM, M., WALKER (éditeur), Madrid, 82-89.

KRUIJF, G-J, SCAAKE, J., (1997), Discerning relevant information in discourses using TFA, in Ruslan, MITKOV, Nicolas, NICOLOV (éditeurs), *Recent advances in Natural Language Processing*, John Benjamins Publishing Company.

KRUIJFF-KORBAYOVA, I., KRUIJFF, G.J.M., (1997), Topic-Focus articulation in DRT, In Proceedings of the 11th Amsterdam Colloguium, December 17-20, 1997.

KRUIJFF-KORBAYOVÁ, Ivana, (1998), *The Dynamic Potential of Topic and Focus: A Praguian Approach to Discourse Representation Theory*, Ph.D Dissertation, Charles University, Prague, June.

KYONGHO, Min, WILSON, William H., (1997), Integrated correction of ill-formed sentences, pp. 369-378 in Advanced Topics in Artificial Intelligence 10th Australian Joint Conference on Artifical Intelligence (AI '97), Lecture Notes in Artificial Intelligence 1342, edited by Abdul Sattar, Berlin: Springer, 1997. ISBN 3-540-63797-4.

LAMEL, L.F., S. K., BENNACEF, H., BONNEAU-MAYNARD, S. ROSSET, and J.-L. GAUVAIN, (1995), Recent developments in spoken language systems for information retrieval, In ESCA ETRW Spoken Dialog Systems, Visgo, Denmark, pages 17-20, May 30-June 2.

LAMEL, Lori, ROSSET, Sophie, GAUVAIN, Jean-Luc, BENNACEF, Samir, (1999), The Limsi Arise system for train travel information, In IEEE International Conference on Acoustics, Speech, and Signal Processing, Phoenix, March.

LANG, Bernard, (1988), Parsing incomplete sentences, COLING'88, Budapest (Hungary), 22 –27, August, vol.1, 365–371.

LANG, Bernard, (1989), A generative view of ill-formed input processing, ATR symposium on basic research for telephone interpretation (ASTI), Kyoto (Japan), 11 – 13 December.

LANG, Bernard, Recognition can be harder than parsing, Computational intelligence, 1994.

LANGACKER, Ronald W., (1987), *Foundations of cognitive grammar*: theoretical prerequisites, Stanford University Press: Stanford.

LANGACKER, Ronald W., (1992), Structural syntax : The view from cognitive grammar, Colloque Internaltional Lucien Tesnière aujourd'hui, Mont-Saint-Aignant, 19 au 21 novembre.

LARREY, P., Compréhension du dialogue oral finalisé par une tâche, Rapport de DEA de l'université Paul Sabatier, 1995.

LASCARIDES, A., COPESTAKE, A., (1998), Pragmatics and Word Meaning, Journal of Linguistics, 34.2, pp387-414, Cambridge University Press. Longer version of paper in Proceedings of COLING 1996.

LAVIE, A., *et al*, (1995), Dialogue processing in a conversational speech translation system, rapport technique, center for machine translation Carnegie Mellon University.

LAVIE, A., *et al*, (1999), The JANUS III translation system : speech to speech translation in multiple domains, C-Star Workshop, Schwetzingen, Germany, September 23-24, 1999. (to appear in Machine translation).

LAZARD, Gilbert, (1994), *L'actance*, Paris : PUF.

LE NY, J. F., (1989), *Sciences cognitives et compréhension du langage*, Paris: PUF.

LEVIN, E, *et al*, (1995), Concept-based spontaneous speech understanding system, Madrid, ESCA Eurospeech.

LEVIN, L., *et al*, (1995), Using context in machine translation of spoken language, in theoretical and methodological issues in machine translation.

LEWIN, I., PULMAN, S. G., (1995), Inference in the Resolution of Ellipsis, Proceedings of ESCA Research Workshop on Spoken Dialogue Systems, March-95 Report CRC-055.

LIFE, A., SALTER, I., TEMEM, J.N., BERNARD, F., ROSSET, S., BENNACEF, S., LAMEL, L., (1996), Data collection for the Mask kiosk: WOz vs prototype system, In International Conference on Speech and Language Processing, pages 1672-1675, Philadelphia, October.

LOPEZ, P., (1999), Analyse d'énoncés oraux pour le dialogue homme-machine à l'aide de grammaires lexicalisées d'arbres, thèse de l'université de Nancy1, Nancy.

LOPEZ, P., (2000), Extended Partial Parsing for Lexicalized Tree Grammars 6th International Workshop on Parsing Technologies (IWPT 2000), Trento, Italy. 23-25 February 2000.

LOPEZ, P., (1998), A LTAG Grammar for parsing oral and incomplete utterances, Proceedings of the 13th European Conference on Artificial Intelligence (ECAI'98), Brighton, UK.23-28 August.

LOPEZ, P., FAY-VARNIER, C., ROUSSANALY, A., (1999), Sous-langage d'application et LTAG : le système EGAL, Proceedings of the 6th Conference on Natural Language Processing (TALN'99), Cargèse, France, July.

LOPEZ, P., (1996), La syntaxe dans les interfaces multimodales intelligentes à composante orale, DEA de l'université Henri Poincaré Nancy 1.

LOPEZ, P., (1999), Représenter et utiliser les contraintes de la langue orale à l'aide d'une grammaire lexicalisée d'arbres adjoints, Proceedings of RECITAL'99, Cargèse, France, July.

LUZZATI, D., (1987), DIALORS: un système de dialogue oral simulé pour une tâche restreinte, XVI èmes JEP, Hammamet, 1987.

MAGERMAN, D., (1994), *Natural language processing as statistical pattern regonition*, Ph.D Dissertation, Stanford University.

MAGERMAN, D., MARCUS, Mitchell P., (1991), Pearl: A Probabilistic Chart Parser, in Proceedings, European ACL, April. Also published in Proceedings, Second International Workshop for Parsing Technologies, February.

MAGERMAN, David M, WEIR, Carl, (1992), Efficiency, Robustness, and Accuracy in Picky Chart Parsing, In Proceedings, ACL Conference, July, 1992.

MANARIS, Bill, HARKREADER, Alan, (1997), SUITE: Speech Understanding Interface Tools and Environments, Proceedings of Tenth International Florida AI Research Symposium (FLAIRS-97), Daytona Beach, FL, pp. 247-252, May.

MANARIS, Bill, HARKREADER, Alan, (1998), SUITE Keys: A Speech Understanding Interface for the Motor-Control Challenged," Proceedings of The Third International ACM Conference on Assistive Technologies (ASSETS '98), Marina del Ray, pp. 108-115, April.

MANNING, Christopher D., CARPENTER, Bob, (1997), Probabilistic Parsing Using Left Corner Language Models, Proceedings of the Fifth International Workshop on Parsing Technologies (IWPT-97), MIT, pp. 147-158.

MARCU, Daniel, (1997), *The Rhetorical Parsing, Summarization, and Generation of Natural Language Texts*, Ph.D Dissertation, Department of Computer Science, University of Toronto, December 1997. Also published as Technical Report CSRG-371, Computer Systems Research Group, University of Toronto.

MARCU, Daniel, (1999), Discourse trees are good indicators of importance in text, In I. Mani and M. Maybury editors, *Advances in Automatic Text Summarization*, pages 123-136, The MIT Press.

MARTINET, André (sous la direction), (1979), *Grammaire fonctionnelle du français*, Paris : Crédif-Didier.

MARTINET, André, (1960), *Eléments de linguistique générale*, Paris : Armon colin.

MARTINET, André, (1985), *Syntaxe générale*, Paris : Armon colin.

MATROUF, A., *et al*, (1987), Système de dialogue orienté par la tâche: une applicaton en avionique, Hammet: XVI èmes JEP.

MAYFIELD, L., *et al*, (1995), Concept based speech translation, Detroit, in proceedings of ICASSP-95.

MAYFIELD, Laura, GAVALDÀ, Marsal, SEO, Y-H., SUHM, Bernhard, WARD, Wayne, WAIBEL, Alex, (1995), Parsing real input in JANUS: a concept based approach, Proceedings of TMI 95.

McKELVIE, D., (1998), SDP - Spoken Dialogue Parser, HCRC Technical Report HCRC/RP-96, May.

MELONI, H., (1996), Prosodie et reconnaissance de la parole, in MELONI, H., (éditeur principal), *Fondements et perspectives en traitement automatique de la parole*, Paris: Aupelf-UREF.

MILLER, Philip, SAG, Ivan A., (1997), French clitic movement without Clitics or Movement, Natural Language and Linguistic Theory, in A. Abeillé et al., *The major syntactic structures of French*, ESSLLI Summerschool, Aix-en-Provence.

MILWARD, D., et al, (1995), Incremental interpretation, Rapport technique EUCCS-WP-1995-1, Center for Cognitive Science, University of Edinburgh.

MINKER, W., BENNACEF, S., (1996), Compréhension et évaluation dans le domaine ATIS. In 21èmes Journées d'études sur la parole, June.

MINKER, W., Waibel, A., Mariani, J., (1999), *Stochastically-based semantic analysis*, Boston: Kluwer Academic Publishers.

MINKER, Wolfgang, GAVALDÀ, Marsal, WAIBEL, Alex, (1999), Hidden Understanding Models for Machine Translation, In European Speech Communication Association Tutorial and Research Workshop on Interactive Dialogue in Multi-Modal Systems (ESCA-1999), Kloster Irsee, Germany, June.

MOESHLER, Jaques, Reboul, A., LUSHER, J.M., JAYEZ, J., (1994), *Langage et pertinence*, Presses Universitaires de Nancy.

MONTEMAGNI, Simonetta, *et al*, (1997), SPARKLE: specification of lexicon structure- Deliverable 2, Final report, November.

MOORE, R., APPELT, D., DOWDING, J., GAWRON, J. M., MORAN, D., (1995), Combining linguistic and statistical knowledge sources in natural-language processing for ATIS, in Proceedings of ARPA Spoken Language Systems Technology Workshop, (Austin, Texas), 22-25 January.

MOORE, R., APPELT, D., DOWDING, J., GAWRON, J. M., MORAN, D., (1995), Combining linguistic and statistical knowledge sources in natural-language processing for ATIS," in Proceedings ARPA Spoken Language Systems Technology Workshop, (Austin, Texas), 22-25 January.

MOREL, Marie-Annick, (1992), Structure hiérarchique de l'énoncé oral, Colloque internaltional Lucien Tesnière aujourd'hui, Mont-Saint-Aignant, 19 au 21 novembre.

MOSNY, Milan, (1996), *Semantic information processing for natural language interfaces to databases*, Thesis of master of science, Simon Fraser University.

MUNK, Markus, (1999), *Shallow statistical parsing for machine translation*, Master Dissertation, Karlsruhe university.

NAKANO, Mikio, (1998), Spoken language analysis based on logical constraint processing, Ph.D. Dissertation, University of Tokyo.

NASH-WEBBER, Bonnie, (1975), The role of semantics in automatic speech understanding, In D., BOBROW, A., COLLINS (éditeurs), *Representation and understanding*: studies in cognitive science, New York: Academic press.

NOMOTO, T., (1997), Effects of grammatical annotation on a topic identification task, in Ruslan Mitkov and Nicolas Nicolov (éditeurs), *Recent advances in Natural Language Processing*, John Benjamins Publishing Company.

NÖTH, E., DE MORI, R., FISCHER; J., GEBHARD, A., HARBECK, S., KOMPE, R., KUHN, R., NIEMANN, H., MAST, M., (1996), An Integrated Model of Acoustics and Language Using Semantic Classification Trees. In *Proc. Int. Conf. on Acoustics, Speech and Signal Processing*, volume 1, pages 419-422, Atlanta.

NOY, N., McGUINESS, D. L., (2001), Ontology development 101: A guide to creating you first ontology, Rapport technique SMI-2001-0880, University of Stanford.

O'SAUGHNESSY, D., (1992), Recognition of hesitation in spontaneous speech, In ICASSP, Pages 521-524.

O'SAUGHNESSY, D., (1993), Analysis and automatic recognition of false starts in spontaneous speech, in ICASSP, Pages II.724-II.727.

O'SAUGHNESSY, D., (1994), Correcting complex false starts in spontaneous speech, in ICASSP, pages I.349-I352.

OCELIKOVA, J., MATROUSEK, V., (1999), Processing of anaphoric and elliptic sentences in a spoken dialog system, EUROSPEECH'99, Budapest Hungary.

PARK, Jun, *et al*, (1999), ETRI speech translation system, C-Star Workshop, Schwetzingen, Germany, September 23-24.

PENSTEIN, ROSE C., (1997), *Robust interactive dialog interpretation*, Ph.D Dissertation, Carnegie Mellon University.

PERENNOU, G., (1996), Compréhension du dialogue oral: Rôle du lexique dans le décodage conceptuel, Séminaire GDR-PRC CHM lexique et communication parlée, octobre.

PIERACCINI, Roberto, LEVIN, Esther, (1993), A Learning Approach to Natural Language Understanding, New Advances and Trends in Speech Recognition and Coding, NATO ASI Series, Springer-Verlag, proceedings of the 1993 NATO ASI Summer School, Bubion, Spain, June-July.

PIERACCINI, R., LEVIN, E., (1992), Stochastic Representation of Semantic Structure for Speech Understanding, Proc. EUROSPEECH 91, September 1991, Genova, Italy, Speech Communication, Vol.11 pp. 283-288.

POLGUERE, A., (1998), La théorie Sens-Texte, Dialangue, Vol. 8-9, Université du Québec à Chicoutimi, pp. 9-30.

POLLARD, C., SAG, I. A., (1994), *Head Driven Phrase Structure Grammar*, CSLI and University of Chicago Press, Stanford, California and Chicago.

POLLOCK Jean-Yves, (1998), *Langage et cognition* : Introduction au programme minimaliste de la grammaire générative, Paris : Presses Universitaires de France.

POTDEVIN, R., (1994), ABC: une architecture logicielle subsymbolique de portée générale pour l'I.A. et la modélisation cognitive, Grenoble: 1ère PCJCSC.

RAJMAN, M., (1996), Approche probabiliste de l'analyse syntaxique in Traitements probabilistes et corpus, revue Traitement Automatique des Langues (TAL), vol. 36, No 1-2, Paris.

RAMM, W., (1997), Discourse constraints on theme selection, in Ruslan MITKOV, Nicolas, NICOLOV, (éditeurs), *Recent advances in Natural Language Processing*, John Benjamins Publishing Company.

RASTIER, F., (1996), Représentation ou interprétation? Une perspective herméneutique sur la médiation sémiotique, in V., RIALLE, D., FISETTE, *Penser l'esprit*, Grenoble : PUG.

RASTIER, F., (1991), *Sémantique et recherches cognitives*, Paris: PUF.

RAYNER, M., CARTER, D. M., (1994), Combining Knowledge Sources to Reorder N-Best Speech Hypothesis Lists, V Digalakis, P Price, ARPA (HLT) Proceedings, Princeton.

REAVES, Benjamin, *et al*, (1999), ATR-Marix : Implementation of a speech translation system, G-Star Workshop, Schwetzingen, Germany, September 23-24.

RENAUD, F., (1982), Présentation d'un modèle linguistique basé sur les langages fonctionnels typés, Intellectia No 6.

REYNIER, E., (1988), *Analyseurs linguistiques pour la compréhension de la parole*, Thèse de l'INPG.

RICHTER, Frank, (2000), *A Mathematical Formalism for Linguistic Theories with an Application in Head-Driven Phrase Structure Grammar*, Doctoral dissertation, University of Tübingen.

RIEGEL, Martin, *et al.*, (1994), *Grammaire méthodique du français*, Paris : Presses universitaires de France.

ROSE, C.P., LAVIE, A., (2001), Balancing Robustness and Efficiency in Unification-augmented Context-Free Parsers for Large Practical Applications, In Robustness in Language and Speech Technology, J. van NOORD, J. C., JUNQUA (éditeurs), ELSNET series, Kluwer Academic Press.

ROUILLARD, J., (1996), Une composante orale de dialogue pour Internet, Rapport de DEA en informatique de l'université Grenoble I.

ROULET, Eddy, AUCHLIN, A., MOESCHLER, J., ROUBATTEL, C., SCHELLING, M., (1987), *L'articulation du discours en français contemporain*, Berne : Peter Lang.

ROUSSEL, David, P. Lopez, (1999), Contribution à l'analyse robuste non déterministe pour les systèmes de dialogue parlé, Proceedings of the 6th Conference on Natural Language Processing (TALN'99), Cargèse, France. July.

SABAH, Gérard, *et al*, (1997), *Machine, langage et dialogue*, Paris : L'Harmattan.

SABAH, Gérard, (1997), Consciousness: a Requirement for Understanding Natural Language, In S. O. NUALLÀIN, P. M., KEVITT, E. M., AOGÀIN (éditeurs), *Two sciences of mind, Advances in Consciousness research*, Amsterdam: John Benjamins, p. 361-392.

SAFIR, Ken, (1998), Symmetry and Unity in the Theory of Anaphora, In Hans, BENNIS, Pierre, PICA, Johan, ROORYCK, (éditeurs), *Atomism and Binding*, Dordrecht : Foris Publications.

SAG, Ivan, WASOW, Thomas, (1997), Syntactic theory: A formal introduction Draft CSLI, 1997.

SANDFORD, E., *et al*, (1997), Une désambiguisation sémantique pour affiner les résultats d'une interrogation d'une base de données textuelle, premières JSF Francil de l'Aupelf.

SARKAR, Anoop, (1997), Separating Dependency from Constituency in a Tree Rewriting System, In Proceedings of the Fifth Meeting on Mathematics of Language, Saarbruecken, Germany, August.

SCHABES, Yves, SHIEBER, Stuart M., (1994), An Alternative Conception of Tree-Adjoining Derivation, Computational Linguistics, volume 20, number 1, pages 91-124.

SHABAN, Marwan, (1993), A minimal GB parser, BU-CS tech report 39-013, Boston.

SHIEBER, Stuart M., (1994), Restricting the Weak-Generative Capacity of Synchronous Tree-Adjoining Grammars, Computational Intelligence 10(4):371-385, November.

SHIEBER, Stuart, PEREIRA, Fernando, DALRYMPLE, Mary, Interactions of Scope and Ellipsis, Linguistics and Philosophy, volume 19, number 5, pages 527-552, October 1996.

SHOPEN, Timothy (éditeur), (1985), *Language Typology and Syntactic Description:* Grammatical Categories and the Lexicon, Cambridge: Cambridge University Press.

SHÜTZE, Carson, (1996), The Empirical Base of Linguistics: Grammaticality Judgments and Linguistic Methodology, University of Chicago Press.

SIU, Man-hung, (1998), *Linear local lexical structure in spontaneous speech language modeling*, Ph.D. Dissertation, Boston University.

SOWA, J., (1982), Using a lexicon of canonical graphs in a semantic interpreter, in WALTON EVENS, M., Relational model of the lexicon: representing knowledge in semantic networks, Cambridge: Cambridge university press.

SPERBER, Dan, WILSON, Deirdre, (1986), *Relevance: Communication and cognition*, Oxford: Blackwell.

SPRIET T., F. BECHET, EL-BEZE M., C. de LOUPY, L. KHOURI, (1996), Traitement Automatique des Mots Inconnus, TALN 96, Marseille France, 22-24 mai.

SPRIET, T, El-Bèze M, (1997), Introduction of Rules into a Stochastic Approach for Language Modelling, NATO ASI Summer school, NATO Book.

SRINIVAS, B., (1997b), Performance Evaluation of Supertagging for Partial Parsing, Proceedings of Fifth International Workshop on Parsing Technology, Boston, USA, September.

STEMMER, Georg, NÖTH, E., NIEMANN, H., (2000), The Utility of Semantic-Pragmatic Information and Dialogue-State for Speech Recognition in Spoken Dialogue Systems. In Karel Pala Petr Sojka, Ivan Kopecek, editor, *Proc. of the Third Workshop on Text, Speech, Dialogue,- TSD 2000*, volume 1902 of *Lecture Notes in Artificial Intelligence*, pages 439-444, Berlin, September 2000. Springer-Verlag.

STOLCKE, A., SHRIBERG, E., (1996), Statistical language modeling for speech disfluencies, Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing, 405-409, Atlanta, GA.

STOLCKE, A., Segal, J., (1994), Precise N-gram Probabilities from Stochastic Context-free Grammars. Proc. ACL, 74-79, Las Cruces, NM.

STOLCKE, A., (1997), Linguistic Knowledge and Empirical Methods in Speech Recognition. AI Magazine 18(4): Winter, 13-24.

STOLCKE, A., (1997), Modeling Linguistic Segment and Turn Boundaries for N-best Rescoring of Spontaneous Speech. Proc EUROSPEECH, 2779-2782, Rhodes, Greece.

STRUBE De LIMA, Vera Lucia, (1990), *Contribution à l'étude du traitement des erreurs au niveau lexico-syntaxique dans un texte écrit en français*, Thèse de doctorat de l'université Grenoble 1.

STURT, P., *et al*, (1995), Incrementality and monotonicity in syntactic parsing, In Incremental interpretation, Edinburgh: Center for cognitive science-university of Edinburgh.

STURT, Patrick, (1997), *Syntactic reanalysis in human language processing*, Ph.D. Dissertation, University of Edinburgh.

SUHM, Bernhard, (1998), *Multimodal interactive error recovery for non-conversational speech user interfaces*, Ph.D. Dissertation, University of Karlsruhe.

SUN, Jiping, TOGNERI, Roberto, DENG, Li, (2000), A robust speech understanding system using conceptual relational grammar, In the proceedings of the 6th international conference on spoken language processing (ICSLP'00), Beijing, China.

TAKEHITO, Utsuro, MATSUMOTO, Yuji, NAGAO, Makoto, (1993), Verbal Case Frame Acquisition from Bilingual Corpora, Proceedings of the 13th International Joint Conference on Artificial Intelligence (IJCAI-93), pp.1150-1156, August.

TAPANAINEN, Pasi, JÄRVINEN, Timo, (1994), Syntactic Analysis of a Natural Language Using Linguistic Rules and Corpus-Based Patterns, In the proceedings of the Fifteenth International Conference on Computational Linguistics (COLING'94). Vol. I, pages 629-634. Kyoto, Japan.

TUE VO, Minh, (1998), *Framework and toolkit for the construction of multimodal learning interfaces*, Ph.D. Dissertation, Carnegie Mellon University.

VAUFREYDAZ, Dominique, BERGAMINI, Carole, SERIGNAT, Jean-François, AKBAR, Mohamad, (2000), A New Methodology for Speech Corpora Definition from Internet Documents LREC`2000, 2nd International Conference on Language Ressources and Evaluation, Athens, Greece, 3 vol. p.423-426.

VERGNE, Jacques, GIGUET, Emmanuel, (1998), Regards Théoriques sur le "Tagging", In proceedings of the fifth annual conference Le Traitement Automatique des Langues Naturelles (TALN 1998), Paris, France, June 10-12.

VERSPOOR, C. M., (1997), *Contextually-dependent lexical semantics*, Ph.D. Dissertation, University of Edinburgh.

VIJAY-SHANKER, K., WEIR, David, (1993), The use of shared-forests in Tree Adjoining Grammar Parsing, in *Proceedings of the 6th Conference of the European Chapter of the Association for Computational Linguistics*, 384-393.

VIJAY-SHANKER, K., WEIR, David, (1995), Parsing D-Tree Grammars, in Proceedings of the International Workshop on Parsing Technologies.

VILLASEÑOR-PINEDA, Louis, (1999), *Contribution à l'apprentissage dans le dialogue homme-machine*, Thèse de Doctorat de l'université Joseph Fourrier-Grenoble 1.

WAIBEL, A., *et al*, (1996), JANUS II: Translation of spontaneous speech conversatonal speech, in proceedings of ICASP-96, Atlanta, May.

WANG, Yeyi, (1998), *Grammar interface and machine translation*, Ph.D. Dissertation, Carnegie Mellon University, Pittsburgh.

WEBBER, Bonnie, (2000), Computational Perspectives on Discourse and Dialogue. In Deborah SCHIFFRIN, Deborah TANNEN, Heidi, HAMILTON (éditeurs), *The Handbook of Discourse Analysis*, Blackwell Publishers Ltd.

WEBBER, Bonnie, JOSHI, Aravind, (1998), Anchoring a Lexicalized Tree-Adjoining Grammar for Discourse, ACL/COLING Workshop on Discourse Relations and Discourse Markers, Montreal, Canada, 15 August 1998.

WEBBER, Bonnie, KNOTT, Alistair, JOSHI, Aravind, (1999), Multiple Discourse Connectives in a Lexicalized Grammar for Discourse, Third International Workshop on Computational Semantics, Tilburg, The Netherlands, January-1999.

WEHRLI, E., (1997), *L'analyse syntaxique des langues naturelles*: problèmes et méthodes, Paris: Masson.

WEIR, David, (1992), Linear context-free rewriting systems and deterministic tree-walking transducers, in Proceedings of the 30th Annual Meeting of the Association for Computational Linguistics.

WHITE, Christopher, (1995), Converting Context Free Grammars to Constraint Dependency Grammars, Masters Thesis, Purdue University, August.

YANG, Li, ZHANG, Tong, LEVINSON, Stephen E., (2000), Word concept model: a knowledge representational for dialogue agents, In the proceedings of the 6th international conference on spoken language processing (ICSLP'00), Beijing, China.

YASUHARU, Den, HARUKI, Yuu, ISHIZAKI, Masato, (1997), A Corpus-Based Analysis of Speech Repairs in Japanese, CPL 97, San francisco.

ZECHNER Klaus, WAIBEL Alex, (1998), Automatic construction of frame representations for spontaneous speech in unrestricted domains, COLING-ACL, Montréal.

ZECHNER, Klaus, (1997), *Building Chunk Level Representations For Spontaneous speech in Unrestricted Domains*: The CHUNKY System and its application to reranking Nbest Lists of speech recogniser, Thèse de Master, Carnegie Mellon University.

ZECHNER, Klaus, (2001), *Automatic Summarization of Spoken Dialogues in Unrestricted Domains*, Ph.D Dissertation, Carnegie Mellon University.

ZEILLIGER, J., *et al*, (1997), Vers une méthodologie d'évaluation qualitative des systèmes de compréhension et de dialogue oral Homme-Machine. In Actes des Journées Scientifiques et Techniques (JST'97), Avignon. Aupelf-Uref.

ZWEIGENBAUM, P., *et al*, (1997), Construction d'une représentation sémantique en graphes conceptuels à partir d'une analyse LFG, Grenoble: 4ème conférence annuelle sur le traitement automatique du langage naturel.

Publications personnelles

ANTOINE, Jean-Yves, BOUSQUET-VERNHETTES, Caroline, GOULIAN, Jérôme, KURDI, Mohamed-Zakaria, Rosset, Sophie, VIGOUROUX, Nadine, VILLANEAU, Jeanne, (2002), Predictive and objective evaluation of speech understanding: the "challenge" evaluation campaign of the I3 speech workgroup of the French CNRS, Third International Conference on Language Ressources and Evaluation LREC02, Las Palmas.

KURDI, Mohamed-Zakaria, (1996), *Perception phonémique et patterns linguistiques* : étude psycholinguistique du traitement lexical, Mémoire de DES en linguistique, Université d'Alep, (118 pages).

KURDI, Mohamed-Zakaria, (1997), *Coopération Syntaxe-prosodie dans la compréhension de la parole : protocole expérimental*, Mémoire de Maîtrise en sciences du langage, Université Lyon II Lumière, (58 pages).

KURDI, Mohamed-Zakaria, (1998), *Reconnaissance de la parole par concept : vers un analyseur robuste des dialogues oraux spontanés*, Rapport de DEA en sciences cognitives, Institut National Polytechnique de Grenoble, (44 pages).

KURDI, Mohamed-Zakaria, (1999), A Chunk based partial parsing strategy for reranking and normalizing Nbest lists of a speech recognizer, ESSLLI'99, Utrecht, Netherlands.

KURDI, Mohamed-Zakaria, (2000), A semantic based approach to spoken dialogue understanding, in D-N Christodoulakis (editor), Natural Language Processing - NLP'00, Lecture notes in Artificial Intelligence 1835, Berlin: Springer.

KURDI, Mohamed-Zakaria, (2000), *Expertise pour la conception d'un système de classification d'e-mails*, Technical report CLIPS-IMAG, September-2000, (101 pages).

KURDI, Mohamed-Zakaria, (2000), La grammaire sémantique d'unification d'arbres : un formalisme pour le l'analyse de la parole spontanée, 7th Conference on Automatic Natural Language Processing TALN'00, Lausanne Swizerland, October 15-18.

KURDI, Mohamed-Zakaria, (2000), The Semantic Tree Unification Grammar: A new formalism for Spoken language parsing, 6th International Conference On Spoken Language Processing ICSLP'00, Beijing China, October 16 - 20.

KURDI, Mohamed-Zakaria, (2000), Une approche intégrée pour la normalization des extragrammaticalités de la parole spontanée, 7th Conference on Automatic Natural Language Processing TALN'00, Lausanne Swizerland, October 15-18.

KURDI, Mohamed-Zakaria, (2001), A spoken language understanding approach which combines the parsing robustness with the interpretation deepness, in the proceedings of the International Conference on Artificial Intelligence IC-AI01, Las Vegas, USA, June 25 - 28.

KURDI, Mohamed-Zakaria, CAELEN, Jean, (1999), SAFIR : un système de compréhension automatique des dialogues oraux spontanés, Journée des jeunes chercheurs en sciences cognitives, Bordeaux.

KURDI, Mohamed-Zakaria, AHAFHAF, Mohamed, (2002), Toward an objective and generic Method for Spoken Language Understanding Systems Evaluation: an extension of the DCR method, Third International Conference on Language Ressources and Evaluation LREC02, Las Palmas.

ROUSSEL, Davide, KURDI Mohamed-Zakaria, CAELEN, Jean, (1999), Normalisation des extragrammaticalités, supertagging et analyse partielle pour le traitement de la parole. Workshop Méthodes hybrides taln / talp pour le traitement robuste du langage, Cargèse, 11-17 juillet-1999.

Annexes

1 Annexe1 : Extraits des corpus utilisés

1.1 Le corpus de réservation hôtelière

sh11.l

H= Hôtel des Tilleuls. Bonsoir.

C= Bonjour Monsieur. Vous reste-t-il une chambre pour une personne pour la nuit de dimanche prochain ?

H= Dimanche prochain. Pour la nuit du 15 au 16. Il me reste des chambres à 250 francs, à 350 francs et à 400 francs.

C= Oui. Mais, quelles sont les différences entre vos chambres ?

H= Les chambres à 250 francs sont beaucoup plus petites. Elles donnent du côté du boulevard, donc elles sont un peu plus bruyantes. Par contre, les chambres à 350 et 400 francs sont très bien. Celles à 350 sont avec douches et à 400 avec bain.

C= Celles à 350 sont bien ?

H= Tout à fait, Madame, ce sont des chambres calmes, agréables. Toutes nos chambres ont été rénovées, il n'y a pas longtemps. Elles sont très bien.

C= D'accord. Vous êtes bien situé à proximité de la Place Jules Ferry, n'est-ce pas ?

H= Tout à fait, Madame. Nous sommes dans la rue qui fait l'angle avec la place.

Eh bien, je vous réserve donc une chambre pour dimanche. Alors nous disons une chambre pour une personne, pour la nuit du 15 au 16 au nom de ?

C= Madame Victor, comme le prénom.

H= Savez-vous vers quelle heure vous arriverez ?

C= Vers 16 heures. Dites-moi, le lendemain, est-ce que je pourrais

vous confier mes bagages en toute sécurité pour la journée ?

Je ne reprends le train que vers 17 heures.

H= Il n'y a aucun problème, vous pouvez être tranquille. Est-ce que vous pouvez me confirmer par fax votre réservation ?

C= Ah non, Monsieur, je suis désolée. Je suis retraitée, je n'ai pas de fax.

H= Dans ce cas, ce n'est pas grave. C'est surtout important quand il y a beaucoup de monde, mais là c'est calme. Donc, c'est entendu à dimanche.

C= Je vous remercie. Au revoir Monsieur.

1.2 Extrait du corpus Nespole

Fichier de transcription du scénario 2 dialogue/ fre002co2.wav»

«C : pour client» «A : pour Agent»

Client : B bonjou ::::::r, c'est l 'agence e APT du Trentino?

Agent : Ah oui, bonjou ::r bonjou : :r Monsieur

Client : Oui bonjou : r, e oui, j'appelle de e / j'appelle de France

A : (m)

C : pour organiser e un voyage avec ma femme donc nous sommes en couple°

A : (m) (m)

C : au mois de septembre et nous aime/ nous aimerions e savoir e ce que l'on peut fai : re e si vous proposez des pakages tout compris

A : : Oui ?

C : : Pour éventuellement e avoir des soins thermaux :/ ::/ e et puis faire des excursions :/ :/ et éventuellement assister à quelques festivals ,

A : : (m) (m)

C : : dans la région

A : oui

C : donc, est ce qu'il existe des cho : ses e ?

A : vous connaissez déjà la région monsieur ou : c'est la première fois que vous venez : (h) ici ?

C : Alors e oui c'est la première (h) fois e c'est la première fois que je viens (h) oui

A : Vous avez des préféren : ces sur certaines localités ?

C : non ** jamais

A : de notre région ?

C : Non e on a pas de préféren : ce sur les localités par contre on aimerait , au niveau des activités pouvoir faire un peu de e de kaya : ck (h) de randonnées : et puis éventuellement assister à des : des festivals ou des cho : ses culturelles qui sont dans e (h) dans (h) cette région là

A : e : oui

C : donc, qu'est ce que vous pourriez me conseiller ?

A : e : en qu'elle période ? en été : ?

C : Alors ça serait pour le mois de septembre°

A : Ah (h) le mois de septembre° , oui

C : (m) (m)

A : alors e (m) (m) pour les localités therma : les° je peux vous avoir°

C : oui ?

A : une offre° e : dans la vil**le de e Levico , Te : rme (h) c'est une localité thermale° (h) et ils : offrent des forf/

C : d'accord co / quel est le nom de la ville ?

A : Levico , , Te : rme (h)

C : d'accord Levico

A : ou : i c'est à côté d'un lac° (h) et : (h) ils : offrent

C : d'accord

A : des forfaits : / : / pour les adu : ltes° (h) et : qui prévoit une semaine relaxante avec des bains° thermaux et des massages ** et : / : / e (m) et des traitements pour les visa : ges & alors sim**ple

C : d'accord

A : et relax on peut dire e (h) et puis si vous voulez

C : très bien

A : les traitements e: (m) sont prévus le matin° et l'après-midi si vous voulez vous pouvez fai : re e des excursions e : à pied :

C : Oh: (h)

A : ou en cheva : l il y a : aussi des manèges là : et : e (m) (h) et vous pouvez choisir aussi

C : d'accord

A : des excursions guidées : par des guides de monta : gne° qui vous illu**/

C : d'accord est-ce que les prix : / le prix que vous m'avez : donnés la :

A : oui ?

C : sont inclus aussi : les l'organisation des excursions : /

A : Alo : rs l'/ les :

C : guidées ou alors ce sont seulement : e : /

A : oui les : / : excursions guidées sont gratuites° , elles sont organisées par le syn**dicat

C : d'accord

A : d'initiative loca : le° et :

C : ok :

A : l’ e les excursions à cheval e : (m) dans les manèges non , e ne sont pas comprises e : (m) en tout cas

C : d’accord

Ales excursions guidées oui e : (m) elles sont déjà payées par le syndicat d'initiative il faut seulement s'inscri : re /

C:ok :

A : et quand vous arrivez : il n'y a pas de problème pour le mois de septembre° il y ’a pas e

beaucoup de mon**de et alo : r , e il faut pas réserver en avance° (h)

C : et qu’est ce qui est : vraiment inclus dans dans le prix alors (h) de e de cette offre ?

A : Alors il y a inclus l’hébergement /

C : il y a l’hôtel /

A : l’hôte : l°,

C : oui :

A : il y a : e (m) la demi-pension° , et :

C : d’accord

A : et puis les les prix e sont ce sont pour personne° , et : à personne ,

C : ok

A : et : la chambre est à deux lits

C : par personne

A : il y a : la visite médi / médicale° , au bout de la semai : ne° et puis il y a des bains : thermaux : , et puis il y a des massa : ges ,

C : Ok

A : trois massages° ** et trois traitements pour le visa : ge° , personnalisés

C : Donc tout : cela est inclus dans le/ dans le prix ?

A : oui , bien sûr oui

C : § c’est génial

A : e : le : voyage c’est pas inclus : et : e et les : excursions°

C : bien sûr

A : ne sont pas inclus mais elles sont gratuites alo : rs e : /

C : Alors / a propos du/ du : voya : ge ?

A : oui : ?

C : justement, a propos du voyage comment faudrait accéder : e dans cette ville ? e : moi e je viendrai depuis la France donc je vais venir en train jusqu'à Milan°

A : (m) (m)

C : mais après depuis Milan comment faut-il faire pour :

A : Alors si vous venez en /

C : pour venir ?

A : Alors si vous venez en train à milan e : vous pouvez continuer en train° e de Milan° sur la ligne Milan° e : Brennero° et : vous descendez à la ville de Trento°

C : d'accord

A : e : ça fait deux/ deux heures et moitié

C : (m) (m)

A : à peu prés trois heures si vous devez changer à Vérone° et :

C : d'accord

A : et puis à Tre : nto il faut que vous preniez l'autobus un autocar il y a des lignes régionales° et : (m) la gare de l'autocar c'est à côté de la ga : re , de chemin de fe : r , vous pouvez la rejoindre

C : d'accord

A : à pied c'est très : / très facile et : vous pouvez trouver des autocars : ici en deux heu : res e : au maximu**ms rejoindre toutes les localités de la région alors pour aller ici à Le : vico e : e : / : / , une demi heu : re, quar**ante minutes

C : Ok:

A : d'autocar

C : d'accord

A : et il y a un autoca : r e : ils sont très souvent, e vous pouvez les trouver tout les jou : rs il y en a : / e il y a beaucoup de courses c'est pas : e c'est pas difficile /

C : Est-ce qu'il y a-les : / les horaires des autocars sur vos : e pages web ?

A ; Non non je le regre**tte nous ne l'avons pas

C : d'accord

A : sur la page web mais en tout cas

C : d’accord

A : e : ça commence ** autobus pour rejoindre Levico c’est très : facile° (h) e : il y en a aussi plusieu : rs (h) dans une heure , et : e : (m) c’est pas difficile e : en tout cas

C : d’accord

A : vous trouverez aussi le train° qui continue pour Levico°, si vous ne trouvez pas l’autocar , vous trouverez : le train° , il y a /

C : d’accord très bien

A : a un petit train local qui : /

C : il y a aussi un train qui va là bas d’acco : rd

A : oui : oui il y a aussi

C : d’acco : rd

A : une petite gare là bas alors c’est très facile° (h) et : si vous voulez choisir le taxi il y a aussi le service taxi mais : il est : /

C: d’acco: rd ,

A : plus che : r alors /

C : c’est plus cher /

A : (m) (m)

C : Bien sûr d’accord ,

A : se sont un peu près trente kilomètres/

C : et alo : rs si non au niveau/, ok : et au niveau culturel alors ? est ce qu’il y a des événements moi je pense partir (h) autour du cinq° septembre début septembre est ce qu’il y a des des cho : ses qui se passent dans la régions des festiva : ls des : (h) ?

A : (m) (m) oui pou : r

C : pour le soir éventuellement

A : pou : r septembre° il y a beaucoup des : des occasions liées surtout : aux fêtes° , , e : (m)(m) aux fê : tes° , e : (m)(m) populai : res° et : au fê : tes°,

C : (m) (m)

A : de la campagne° alo : rs (h)

C : d’accord

A : liées : surtou : t aux vendanges ° , on dit comme ça (=) a : e

C : Oui les vendanges

A : les fêtes de raisins oui à la vendange fêtes des pommes des fruits en général° il y a : / non les fêtes de marrons se sont plus tar/ plus tard il y a aussi des/ des fêtes liées au animaux : oui : Toutes les activités

C : (m) (m)

A : de la montagne de la et de la compagne° au Valsugana

C : d'accord

A : par exemple° à la fin de septembre°, à partir de : e vingt trois vingt quatre septembre° il y a une fête populai : re liée aux activités e : : (m) de la montagne° , aux activités : e : (m) pour le transfèrement des animaux de l' : : e haute montagne° à la moitié montagne° à la fin de l'été : , elle s'appelle sa**

C : Et à quoi cela consiste ?

A : e : : ils fo**nt :/

C : il y a des spectacles° e : : des : : ?

A : il y a des spectacles folkloriques° il y a aussi des : e des évocations des anciens métiers : il y a : : : (m)

C : d'accord

A : des travaux artisanaux et tout ça il y a aussi de la musique° (h) typique° il y a : des festins gastronomiques° et : : ce sont deux jours/

C : (m)

A : deux journées : : complètes° , (h) nous n'avons

C : d'accord

A : pas encore le : : e : : le programme des : : festives mais en tout cas : :

C :Bien sûr

A : il s'agit de : : e de tout ce/ de tout ce ça , et : : : : (m) et puis , e : : : :

C : Ok bon et ben merci /

A : au bout de septembre/ , Oui il y a aussi

C : oui ?

A : des évocations historiques° au bout de septembre au Valsuga : :na , par **le lieu de la Brenta

C : ok : :

A : ce sont des jeux° anciens sur le fleuve qui s'appelle Brenta, c'est le fleuve de la Valsugana

C :D'accord

A :et ce sont des jeux médiévaux

C : Alors j'avais (h) / j'avais une dernière question au niveau des activités, en plus des excursions est ce qu'on

peut faire du kayak ou de la planche à voile sur le lac (h)?

A : eh : : oui

C : ou e /

A : sur le lac du Levico on peut faire le kayak° mais pas de la planche à voile°

C : (m) (m)

A : il faut : : rejoindre /

C : pas la planche d'accord

A : non il faut rejoindre le : : e le lac de Caldonazzo que c'est tous prés du : lac Levico , et : : : : (m)

C : (m) (m)

A : et là : : vous pouvez trouver : : trois centres° (h) de planche à voile° (h) oui sur le lac de /

C : d'accord est ce que l'eau : / l'eau est chaude ? quelle est la températu : :re ?/

A : Oui en septembre /

C : environ en septembre(=)?

A : l'eau : : est vraiment chaude pas moins de : : e

C : d'accord

A : dix degrés je/ je crois oui en septembre elle est chaud**e , oui ça va

C : ok : :

A : et vous pouvez louer aussi directement l'équipement sur place (h)

C : D'accord , et alors en ce qui concerne le e : : pak/ , le e : : l'offre que vous m'avez proposée est ce que je dois la réserver auprès : : de vous ? ou auprès de l'hôtel directement?

A : Vous pouvez répéter s'il vous plaît ? e

C : e en ce qui concerne l'offre l'offre

A : Ah : : oui ?

C : donc de d'une semaine

A : oui ?

C : Est-ce que je dois la réserver auprès de votre office APT ?

A : Non

C : ou alors auprès de l'hôtel directement ?

A : Non , il faut que vous la réserviez directement à l'hôtel /*/ nous ne pouvons pas faire des réservations non

C : D'accord

A : Vous avez trouvé les : : e / l'adresse de l'agence° (h) ?, e : : sur la page que e n**ous vous avons envoyé vous l'avez trouvé ?

C : Ouais

A : (m) (m)

C : e : : oui : : j/ j'ai la/ j'ai les pages web oui

A : sur la page web /

C : Est-ce que / (h) est ce que vous pouvez me donner/ oui : ?

A : Oui, vous trouverez l'adresse là-bas sur la page , web oui : : ?

C : Ok : : ok ben je vous remercie alors,

A : Ok

C : merci pour ces informations.

A : Merci à vous au revoir bon séjour.

C : Au revoir.

A : Au revoir.

C : Au revoir.

1.3 Extrait du Trains Corpus

Voici un dialogue extrait du *Trains Corpus*.

Dialogue : d93-23.2

Number of utterances files: 98

Length of dialogue: 382.951348

Estimated number of turns: 70

```
utt1  : : s :  hi can I help you
utt2  : : u :  uh <sil> yeah <sil> well <sil> uh <sil> okay <noise> <sil> uh
utt3  : :     okay <sil> first thing I want to do is <sil> move <sil>
            two engines <sil> and <sil> two boxcars <sil> to Corning <sil>
            from Elmira
utt4  : : s :  okay <sil> so I'll send engine E two first and then <sil>
            engine E three right + after it +
utt5  : : u :  + right +
utt6  : :     right
utt7  : : s :  okay
utt8  : : u :  and <sil> then <sil> oh what time is it
utt9  : : s :  it's midnight
utt10 : : u :  oh it's midnight <sil> okay <brth> uh <sil> okay <sil> then
utt11 : :     I want to <sil> do two things at once then I want to send <sil>
            two tankers <sil> from Corning <sil> to Elmira
utt12 : : s :  the <sil> tankers have to have an engine attached to them
utt13 : : u :  oh they do
utt14 : : s :  yeah but you can put <sil> more than one tanker <sil> on an engine
utt15 : : u :  oh I can + <sil> can I + put more than one boxcar on an engine
utt16 : : s :  + yeah +
utt17 : :     uh you can put <sil> wait <sil> you can put <sil> three <sil>
               loaded boxcars <sil> or tanker cars <sil> on an engine <sil> and
<sil>        any number of unloaded cars
utt18 : : u :  oh <sil> oh okay <sil> oh okay <sil> well <sil>
            okay then I want to uh <sil> then I want to <sil>
```

take one of the engines

utt19 : : s : mm-hm

utt20 : : u : and put <sil> two tankers <sil> on the engine <sil>

from Corning to Elmira

utt21 : : s : in addition <sil> to <sil> the boxcar

utt22 : : u : no no no leave the boxcars at Corning

utt23 : : s : okay

utt24 : : u : take one engine and <sil> and and two <sil> and two tankers and <sil> put it take it to Elmira

utt25 : : s : okay so I'll send E two <sil> then

utt26 : : u : yeah

utt27 : : s : alright

utt28 : : u : okay <sil> and then <sil> how long does it take to get to <sil> from Corning to Dansville

utt29 : : s : one hour

utt30 : : u : and how long does it take to get from Corning to Bath

utt31 : : s : two hours

utt32 : : u : okay <brth> then <sil> I want to send <sil> um <sil> one engine

utt33 : : and <sil> one <sil> and and and <sil> two boxcars from Corning to Dansville

utt34 : : s : okay so <sil> E three has already got <sil>

a boxcar on it and it's at Corning <sil> shall I

utt35 : : u : go- it's already got <sil> two boxcars on it

utt36 : : s : um <sil> there's one boxcar that's just sitting there left over from + E + <sil> + E two +

utt37 : : u : + right +

utt38 : : + now put that + + on top + of it

utt39 : : s : + okay +

utt40 : : u : put + that on + top of it

utt41 : : s : + alright +

utt42 : okay + so + <sil> E three and two boxcars go to Dansville from Corning okay

utt43 : u: + and then s- +

utt44 : right <sil> and then <sil> right go go from Corning to Dansville

utt45 : s: yeah you're right <sil> + sorry +

utt46 : u: + right +

utt47 : and then <sil> okay then <sil> I want <sil>
to do two things at once again

utt48 : I want to <sil> load <sil> the two tankers <sil>
at Elmira with orange juice

utt49 : s: okay

utt50 : did you load the boxcars with oranges at Corning

utt51 : u: no

utt52 : s: okay

utt53 : u: okay are they loaded <sil> with oranges

utt54 : s: I mean with

utt55 : u: with orange + juice +

utt56 : s: + yes + <sil> + they + are

utt57 : u: + right +

utt58 : okay <sil> and then I want to send that <sil>
and then I want to send those two tankers back to Corning

utt59 : s: okay

utt60 : fine

utt61 : u: and then <sil> I want to pick up another boxcar at Dansville <sil> with <sil> the <sil> two <sil> boxcars that are already on engine three

utt62 : s: okay

utt63 : u: and send <sil> all three of those <sil> to Avon

utt64 : s: alright so <sil> three <sil> empty boxcars with engine E three <sil> go to Avon

utt65 : u: right

utt66 : s: alright

utt67 : u: okay <sil> and then <sil> I want to <sil> load <sil>
all those boxcars with bananas

utt68 : s: alright

utt69 : u: + and +

utt70 : s: + okay +

utt71 : u: and then <sil> I want to send <sil> all those boxcars of bananas <sil> to Bath <sil> and also <sil> I want to send the tankers <sil> with orange juice to Bath

utt72 : s: okay

utt73 : u: and then we're done

utt74 : s: alright did we get there what time did you have to arrive at Bath

utt75 : u: um by noon

utt76 : s: by noon let's see if we made it

utt77 : so <sil> you sent <sil> engine <sil> engine E two <sil> and engine <sil> E three <sil> to <sil> Corning

utt78 : u: mm

utt79 : s: each with a boxcar <sil> and that <sil> took <sil> two hours <sil> then

utt80 : you s- at the same time you sent <sil> um <sil> E two <sil>
uh with <sil> with three tankers <sil> to <sil> Elmira

utt81 : so that takes two hours and E three is going to Dansville <sil>
with the two boxcars and that takes one hour

utt82 : so let's see

utt83 : two a.m. <sil> Corning

utt84 : and okay <brth> so four a.m. <sil> uh <sil> E <sil> two <sil>
is at <sil> Elmira

utt85 : three a.m. <sil> E <sil> three <sil> is at Dansville

utt86 : um <sil> it takes six hours <sil> to get from Dansville <sil>
to Avon <sil> so that's <sil>
i- it takes three hours so that's um six a.m.

utt87 : then you load them there so that's seven a.m.

utt88 : and send them to Bath <sil> that takes four hours

utt89 : so that's <sil> eleven <sil> a.m. so that <sil>
much of it works out now let's see what engine E two is doing

utt90 : um <sil> E two gets <sil> to Elmira at four a.m.

utt91 : it <sil> is <sil>
lo- the tankers are loaded with orange juice so that's <sil> five a.m.

utt92 : uh <sil> it goes <sil> back <sil> to Corning <sil>
that's two hours so it's seven a.m.

utt93 : and <sil> um <sil> you're <sil> taking the orange juice to Bath right

utt94 : u: + mm-hm +

utt95 : s: + so + <sil> yeah that's two hours to Bath so that's <sil>
nine a.m. okay <sil> you said you wanted to get there by

utt96 : u: by noon

utt97 : s: by noon <sil> great

utt98 : u: okay

1.4 Extrait du corpus des meilleures hypothèse de reconnaissance utilisées pour tester Oasis

Voici un extrait du corpus des meilleurs hypothèses de reconnaissance du système Raphael que nous avons utilisé pour testr Oasis :

- [allô,l,hôtel,du,nord],
- [est,ce,qu,il,vous,reste,des,chambres,ordinaires],
- [voilà,c,est,sans,lavabo,je,voudrais,madame,les,catégories],
- [pour,deux],
- [pour,quatre,jours,à,partir,de,demain,quel,est,le,cas,à,if],
- [un,bon,lait,avec,le,petit,déjeuner],
- [je,n,ai,pas,bien,le,choix,enfin,tant,pis,je,la,prends],
- [monsieur,le,clips,avec,ses,combats,mille,c,l,i,p,s,a],
- [oui,mardi,et,le,leur],
- [bonjour,j,aurais,retenu,une,chambre,une,chambre,pour,deux,personnes,ce,serait,pour,demain,pour,demain,soir],
- [oui,non,euh,qu,est,ce,qui,est,possible],
- [bon,attendez,qu,est,ce,que,tu,préfères,bon,bon,d,accord,ça,suffira],
- [allô,madmoiselle,excusez,moi,le,prendre,avec,lavabo,s,il,vous,plaît],
- [oui,nous,avons,quatre,jours,à,paris,enfin,trois,nuits,nous,retenons,la,chambre,pour,trois,nuits,vous,êtes,à,quelle,distance,de,la,gare],
- [merci,merci,bien,et,c,est,calme,la,année,parce,que,ma,femme,a,le,sommeil,léger],
- [parfait,alors,nous,arriverons,demain,par,le,train,de,lyon],
- [non,pas,à,neuf,heures,le ,train,à,dix,huit,pages,mois,jeudi,les,verts,la,ville,le,train,de,année,hier],
- [vous,êtes,tout,excuser,design,en,banque,à,dix,huit,heures],
- [le,schéma,de,en,clips],
- [bonjour,est,il,possible,de,réserver,une,chambre,pour,le,quinze],
- [trois,nuits],
- [j,aimerais,une,chambre,plutôt,grande,avec,télévision,téléphone,baignoire,et,surtout,très,calme],
- [bonjour,je,suis,bien,à,l,hôtel,ibis],
- [bonjour,vous,reste,des,chambres,pour,la,taille,de,du,treize,au,seize,août],
- [c,est,parfait,pouvez,vous,la,réserver,du,treize,au,seize,août,au,nom,de,clips],
- [je,vais,merci,au,revoir],
- [bonjour,auriez,vous,des,chambres,à,mois,de,quatre,cents,francs,s,il,vous,plaît],

- [dans,ce,cas,pouvez,vous,me,réserver,une,de,ces,chambres,pour,la,nuit,du,dix,au,onze,août,s,il,vous,plaît],
- [allô,bonjour,madame,je,voudrais,retenir,une,chambre],
- [eh,bien,ce,serait,pour,deux,nuitées,les,neuf,ce,le,dix,septembre,prochains],
- [pour,une,personne,quels,sont,les,prix,des,chambres,s,il,vous,plaît],
- [dont,alors,c,est,d,accord,pour,les,nuits,de,lundi,neuf,et,de,mardi,dix],
- [monsieur,martin,faut,il,confirmer,par,courrier],
- [donc,de,juin,si,à,lundi,donc,j,arriverai,vers,dix,neuf,heures,trente],
- [bonjour,monsieur,je,voudrais,savoir,ce,que,vous,avez,encore,des,chambres,disponibles,pour,le,quinze],
- [pour,deux,personnes,s,il,vous,plaît],
- [bon,je,voudrais,réserver,chambre,avec,bain,à,carte,sont,en,franc,mais,plutôt,calme,si,c,est,possible],
- [nous,arrivons,en,cinq,d,après,midi,je,ne,s,est,pas,quelle,heure,mais,avant,dix,neuf,heures,tout,ça,sans,eux,la,terre],
- [merci,à,bientôt,donc],
- [non,deux,chambres,chacune,pour,une,personne],
- [oui,c,est,ça],
- [c,est,parfait],
- [au,nom,de,messieurs,jean,et,du,pont],
- [à,bientôt],
- [dans,son,échelle,c,est,bien,l,hôtel,ibis],
- [qu,est,ce,qui,vous,reste,des,chambres,disponibles,pour,trois,nuitées,à,compter,du,mardi,quinze,octobre],
- [de,trois,chambres,une,pour,deux,personnes,et,deux,chambres,pour,une,seule,personne],
- [vos,chambres,sont,à,quel,prix],
- [oui,ça,me,confier],
- [bien,d,accord,ça,ira,dois,je,vous,envoyer,des,adresses],
- [revoir,la,date,août],
- [bonjour,je,cherche,une,chambre,pour,trois,jours,s,il,vous,plaît],
- [a,partir,de,ce,soir,trois,nuit,quelques,idées,chambre,avec,une],
- [est,ce,bien,je,la,prends],
- [chevaux,merci,à,plus,tard],
- [bonjour,madame,je,viens,de,la,part,de,l,office,tourisme,il,paraît,que,vous,avez,encore,des,chambres,libres],

- [eh,bien,je,voudrais,une,chambre,pour,deux,personnes,si,possible,et,surtout,assez,calme,avec,bain],
- [deux,nuits,ce,soir,et,demain,soir],
- [bon,c,est,d,accord,je,vais,rester,ma,femme],
- [très,bien,je,vais,faire,comme,ça,à,tout,de,suite],
- [bonjour,je,suis,issue,martin,j,avais,retenu,une,chambre,pour,une,semaine],
- [c,est,bien,ça],
- [oui,c,est,exact],
- [bon,d,accord,je,repasserai,vers,dix,huit,heures],
- [bonjour,auriez,vous,une,chambre,disponible,pour,dimanche,soir,s,il,vous,plaît],
- [pour,une,nuit,seulement,et,pour,une,personne],
- [une,chambre,avec,bain,plutôt],
- [parfait,vous,me,la,réservez,mais,j,arriverai,tard,au,train,de,vagues,les,heures,qui,arrive,de,lyon,je,crois],
- [monsieur,martin],
- [ah,bon,c,est,bien,et,si,les,revoir],
- [allô,bonjour,je,voudrais,réserver,deux,chambres,pour,trois,jours,s,il,vous,plaît],
- [a,partir,de,mercredi,vingt,au,soir],
- [oui,c,est,ça],
- [vos,chambres,sont,à,quel,prix],
- [les,petits,déjeuners,sont,compris],
- [bonsoir,madmoiselle],
- [ah,bon,eh,bien,écoutez,je,retiens,les,deux,chambres,pour,les,nuits,de,mercredi,vingt,et,je,vais,chercher,autre,chose,pour,les,nuit,suivants],
- [c,est,dire,au,revoir],
- [bonjour,je,voulais,savoir,si,vous,restez,chambre,pour,deux,nuits,à,compter,de,mardi,prochain,s,il,vous,plaît],
- [ah,bon,alors,c,est,d,accord,vous,me,la,réservez,au,nom,de,ce,matin],
- [o,j,arriverais,au,dernier,train,de,nuit,car,à,trois,heures,ce,cours],
- [quinze,douze],
- [entendu,j,ai,bien,compris,la,chambre,trois,cent,neuf,ce,del,code,d,entrée,vingt,zéro,cinq,je,vous,remercie,à,mardi],
- [bonjour,monsieur,je,suis,bien,à,l,hôtel,ibis],
- [j,aurais,voulu,savoir,s,il,faut,laisser,des,chambres,pour,le,quatorze,octobre,juste,une,nuit],

- [j,aurais,une,chambre,pour,une,personne,avec,laquelle,il,toilette,dans,la,chambre,si,possible],
- [oui,avec,douche],
- [bon,alors,très,bien],
- [donc,très,bien,je,voue,merci,au,revoir,monsieur],
- [bonjour,l,hôtel,ibis],
- [je,voudrais,réserver,une,chambre,pour,trois,personnes],
- [bonsoir,je,suis,bien,à,l,hôtel,ibis],
- [je,voudrais,une,chambre,pour,deux,personnes],
- [le,trois,et,le,quatre,janvier],
- [très,bien,chaud,l,air,si,au,vent,de,sud],
- [bonsoir,j,vous,appelle,pour,isoler,une,chambre,pour,mercredi,qui,vient,le,onze],
- [une,seule],
- [au,nom,de,françois,martin],
- [merci,au,revoir,monsieur],
- [bonjour,j,aurais,voulu,une,chambre,pour,le,quatorze],

1.5 Extrait du corpus utilisé pour tester Corrector

Voici un extrait du corpus utilisé pour tester Corrector avec l'analyse morphologique obtenue avec le tagger de Xerox :

- [('uh','+ITJ'),('the','+DET'),('bananas','+NOUN'),('go','+VPRES'),('to','+PREP'),('um','+guessed+ADJ'),('to','+PREP'),('elmira','+guessed+ADJ')],
- [('and','+COORD'),('it','+PRONPERS'),('also','+ADV'),('says','+VPRES'),('that','+COSUB'),('all','+QUANT'),('the','+DET'),('all','+QUANT'),('the','+DET'),('engines','+NOUN'),('except','+VPRES'),('for','+PREP'),('e','+PROP'),('two','+CARD'),('are','+VBPRES'),('going','+VPROG'),('undergoing','+ADJING'),('routine','+ADJ'),('maintenance','+NOUN'),('we','+PRONPERS'),('can','+VAUX'),('only','+ADV'),('use','+VINF'),('engine','+NOUN'),('e','+PROP'),('two','+CARD')],
- [('well','+ADV'),('like','+COSUB'),('I','+PRONPERS'),('said','+VPAST'),('we','+PRONPERS'),('should','+VAUX'),('we','+PRONPERS'),('should','+VAUX'),('do','+VDINF'),('the','+DET'),('bananas','+NOUN'),('first','+ORD')],
- [('um','+guessed+NOUN'),('and','+COORD'),('we','+PRONPERS'),('need','+VPRES'),('three','+CARD'),('boxcars','+NOUN'),('of','+PREP'),('of','+PREP'),('uh','+ITJ'),('bananas','+NOUN'),('so','+COSUB'),('I','+PRONPERS'),('would','+VAUX'),('say','+VINF'),('we','+PRONPERS'),('take','+VINF'),('e','+PROP'),('two','+CARD')],
- [('three','+CARD'),('ok','+ADV'),('so','+ADV'),('so','+COSUB'),('it','+PRONPERS'),('is','+VBPRES'),('six','+CARD'),('hours','+NOUN'),('total','+NOUN'),('to','+INFTO'),('get','+VINF'),('there','+ADV')],
- [('I','+PRONPERS'),('I','+PRONPERS'),('did','+VDPAST'),('not','+NOT'),('say','+VINF'),('nine','+CARD'),('I','+PRONPERS'),('said','+VPAST'),('nine','+CARD'),('a-m','+guessed+NOUN'),('did','+VDPAST'),('not','+NOT'),('I','+PRONPERS')],
- [('right','+ADJ'),('um','+guessed+NOUN'),('but','+COSUB'),('in','+PREP'),('it','+PRONPERS'),('in','+PREP'),('it','+PRONPERS'),('it','+PRONPERS'),('is','+VBPRES'),('midnight','+NOUN'),('now','+ADV'),('too','+ADV'),('right','+ADJ')],
- [('ok','adv'),('so','+COSUB'),('I','+PRONPERS'),('I','+PRONPERS'),('am','+VBPRES'),('I','+PRONPERS'),('am','+VBPRES'),('thinking','+VPROG'),('the','+DET'),('best','+ADJSUP'),('thing','+NOUN'),('to','+INFTO'),('do','+VDINF'),('would','+VAUX'),('be','+VBINF'),('to','+PREP'),('do','+VDPRES'),('the','+DET'),('bananas','+NOUN'),('first','+ORD'),('since','+COSUB'),('they','+PRONPERS'),('are','+VBPRES'),('the','+DET'),('things','+NOUN'),('with','+PREP'),('the','+DET'),('time','+NOUN'),('limit','+NOUN')],

- [('um','+guessed+NOUN'),('so','+ADV'),('let','+VPRES'),('us','+PRONPERS'),('go','+VPRES'),('all','+QUANT'),('the','+DET'),('way','+NOUN'),('to','+INFTO'),('uh','+ITJ'),('to','+PREP'),('to','+PREP'),('avon','+guessed+ADJ'),('through','+PREP'),('dansville','+guessed+ADJ')],
- [('we','+PRONPERS'),('have','+VHPRES'),('two','+CARD'),('two','+CARD'),('bananas','+NOUN')],
- [('at','+PREP'),('the','+DET'),('same','+ADJPRON'),('time','+NOUN'),('we','+PRONPERS'),('are','+VBPRES'),('we','+PRONPERS'),('are','+VBPRES'),('getting','+VPROG'),('an','+DET'),('engine','+NOUN'),('from','+PREP'),('elmira','+guessed+ADJ'),('with','+PREP'),('the','+DET'),('boxcars','+NOUN')],
- [('oranges','+NOUN'),('to','+PREP'),('so','+ADV'),('so','+COSUB'),('we','+PRONPERS'),('have','+VHPRES')],
- [('or','+COORD'),('actually','+ADV'),('no','+ADV'),('I','+PRONPERS'),('am','+VBPRES'),('sorry','+ADJ'),('it','+PRONPERS'),('would','+VAUX'),('be','+VBINF'),('better','+ADJCMP'),('to','+INFTO'),('leave','+VINF'),('leave','+VINF'),('leave','+VINF'),('the','+DET'),('boxcar','+NOUN')],
- [('so','+ADV'),('so','+ADV'),('let','+VPRES'),('us','+PRONPERS'),('see','+VPRES'),('you','+PRONPERS'),('are','+VBPRES'),('gonna','+guessed+NOUN'),('take','+VPRES'),('one','+CARDONE'),('one','+CARDONE'),('engin','+guessed+ADJ')],
- [('and','+COORD'),('it','+PRONPERS'),('is','+VBPRES'),('midnight','+NOUN'),('and','+COORD'),('and','+COORD'),('that','+COSUB'),('so','+ADV'),('takes','+VPRES'),('um','+guessed+ADJ'),('eleven','+CARD'),('hours','+NOUN'),('in','+PREP'),('all','+QUANT')],
- [('ok','+ADV'),('it','+PRONPERS'),('is','+VBPRES'),('it','+PRONPERS'),('is','+VBPRES'),('got','+VPAP'),('three','+CARD'),('boxcars','+NOUN')],
- [('well','+ADV'),('you','+PRONPERS'),('have','+VHPRES'),('got','+VPAP'),('these','+PRON'),('these','+DET'),('three','+CARD'),('boxcars','+NOUN'),('that','+COSUB'),('you','+PRONPERS'),('have','+VHPRES'),('taken','+VPAP'),('from','+PREP'),('dansville','+guessed+ADJ')],
- [('ok','+ADV'),('so','+ADV'),('as','+PREPADVAS'),('at','+PREP'),('five','+CARD'),('a-m','+guessed+NOUN'),('it','+PRONPERS'),('is','+VBPRES'),('got','+VPAP'),('engine','+NOUN'),('two','+CARD'),('has','+VHPRES'),('brought','+VPAP'),('two','+CARD'),('boxcars','+NOUN'),('of','+PREP'),('oranges','+NOUN'),('to','+PREP'),('to','+PREP'),('elmira','+guessed+ADJ'),('they','+PRONPERS'),('are','+VBPRES'),('there','+ADV')],
- [('no','+ADV'),('no','+DET'),('time','+NOUN'),('at','+PREP'),('all','+QUANT')],
- [('ok','+ADV'),('then','+ADV'),('send','+VINF'),('ok','+ADV'),('this','+PRON'),('is','+VBPRES'),('is','+VBPRES'),('should','+VAUX'),('have','+VHINF'),('done','+VDPAP'),('this','+PRON'),('in','+PREP'),('the','+DET'),('beginning','+ADJING'),('engine','+NOUN'),('three','+CARD')],

- [('ok','+ADV'),('it','+PRONPERS'),('is','+VBPRES'),('faster','+ADJCMP'),('to','+INFTO'),('go','+VINF'),('to','+PREP'),('from','+PREP'),('corning','+PARTPRES'),('to','+PREP'),('avon','+guessed+ADJ'),('through','+PREP'),('dansville','+guessed+NOUN'),('that','+PRONREL'),('takes','+VPRES'),('four','+CARD'),('hours','+NOUN')],
- [('that','+PRONREL'),('is','+VBPRES'),('five','+CARD'),('in','+PREP'),('the','+DET'),('morning','+NOUN'),('I','+PRONPERS'),('still','+ADV'),('have','+VHINF'),('plenty','+NOUN'),('of','+PREP'),('time','+NOUN'),('and','+COORD'),('then','+ADV'),('and','+COORD'),('then','+ADV'),('it','+PRONPERS'),('is','+VBPRES'),('four','+CARD'),('hours','+NOUN'),('back','+ADV')]
- [('then','+ADV'),('that','+PRON'),('is','+VBPRES'),('not','+NOT'),('a','+DET'),('problem','+NOUN'),('but','+COSUB'),('I','+PRONPERS'),('do','+VDPRES'),('not','+NOT'),('see','+VINF'),('how','+WADV'),('I','+PRONPERS'),('can','+VAUX'),('get','+VINF'),('the','+DET'),('boxcar','+NOUN'),('of','+PREP'),('bananas','+NOUN'),('and','+COORD'),('the','+DET'),('boxcar','+NOUN'),('of','+PREP'),('oranges','+NOUN'),('to','+PREP'),('bath','+NOUN'),('by','+PREP'),('twelve','+CARD'),('if','+COSUB'),('we','+PRONPERS'),('gotta','+guessed+ADJ'),('go','+NOUN'),('all','+QUANT'),('over','+PREP'),('the','+DET'),('place','+NOUN'),('to','+INFTO'),('pick','+VINF'),('up','+ADV'),('the','+DET'),('boxcars','+NOUN')]
- (('um','+guessed+ADJ'),('the','+DET'),('banana','+NOUN'),('warehouse','+NOUN'),('is','+VBPRES'),('in','+ADV'),('is','+VBPRES'),('in','+PREP'),('avon','prop')],
- [('ok','+ADV'),('now','+ADV'),('we','+PRONPERS'),('still','+ADV'),('have','+VHINF'),('to','+INFTO'),('deal','+VINF'),('with','+PREP'),('now','+ADV'),('we','+PRONPERS'),('still','+ADV'),('have','+VHINF'),('to','+INFTO'),('deal','+VINF'),('with','+PREP'),('our','+DET'),('tanker','+NOUN'),('of','+PREP'),('orange','+ADJ'),('juice','+NOUN')],
- [('instead','adv'),('of','+PREP'),('going','+PARTPRES'),('to','+PREP'),('bath','+NOUN'),('how','+WADV'),('about','+ADV'),('if','+COSUB'),('it','+PRONPERS'),('went','+VPAST'),('to','+PREP'),('corning','+PARTPRES'),('ooh','+ITJ'),('how','+WADV'),('about','+ADV'),('this','+PRON'),('it','+PRONPERS'),('can','+VAUX'),('go','+VINF'),('into','+PREP'),('bath','+NOUN'),('drop','+NOUN'),('off','+PREP'),('the','+DET'),('bananas','+NOUN')],
- [('so','+ADV'),('from','+PREP'),('elmira','+guessed+ADJ'),('to','+PREP'),('corning','+PARTPRES'),('with','+PREP'),('boxcar','+NOUN'),('we','+PRONPERS'),('then','+ADV'),('we','+PRONPERS'),('fill','+VPRES'),('up','+PREP'),('the','+DET'),('boxcar','+NOUN'),('with','+PREP'),('oranges','+NOUN')],
- [('and','+COORD'),('then','+ADV'),('hm','+guessed+ADJ'),('we','+PRONPERS'),('how','+WADV'),('about','+ADV'),('we','+PRONPERS'),('go','+VPRES'),('back','+ADV'),('to','+PREP'),('elmira','+guessed+ADJ')],

- [('but','+COSUB'),('you','+PRONPERS'),('can','+VAUX'),('you','+PRONPERS'),('can','+VAUX'),('carry','+VINF'),('more','+QUANTCMP'),('more','+QUANTCMP'),('than','+COTHAN'),('one','+CARDONE'),('boxcar','+NOUN')],
- [('oh','+NOUN'),('no','+ADV'),('I','+PRONPERS'),('only','+ADV'),('need','+VPRES'),('one','+CARDONE'),('one','+CARDONE'),('boxcar','+NOUN')],
- [('ok','+ADJ'),('so','+COSUB'),('it','+PRONPERS'),('is','+VBPRES'),('so','+ADV'),('engine','+NOUN'),('e','+PROP'),('two','+CARD')],
- [('and','+COORD'),('then','+ADV'),('ok','+ADV'),('now','+ADV'),('we','+PRONPERS'),('only','+ADV'),('have','+VHPRES'),('now','+ADV'),('we','+PRONPERS'),('only','+ADV'),('have','+VHINF'),('two','+CARD'),('we','+PRONPERS'),('have','+VHPRES'),('all','+QUANT'),('three','+CARD'),('open','+ADJ'),('boxcars','+NOUN')],
- [('maybe','+ADV'),('we','+PRONPERS'),('could','+VAUX'),('even','+ADV'),('drop','+VINF'),('a','+DET'),('boxcar','+NOUN'),('just','+ADV'),('leave','+VINF'),('leave','+VINF'),('leave','+VINF'),('it','+PRONPERS'),('and','+COORD'),('go','+VPRES'),('to','+INFTO'),('go','+VINF'),('to','+PREP'),('avon','+guessed+ADJ'),('fill','+NOUN'),('up','+ADV'),('with','+PREP'),('the','+DET'),('bananas','+NOUN'),('and','+COORD'),('come','+VINF'),('back','+ADV'),('to','+PREP'),('dansville','+guessed+ADJ')],
- [('you','+PRONPERS'),('are','+VBPRES'),('trying','+VPROG'),('to','+INFTO'),('get','+VINF'),('you','+PRONPERS'),('are','+VBPRES'),('trying','+VPROG'),('to','+INFTO'),('get','+VINF'),('all','+QUANT'),('of','+PREP'),('the','+DET'),('oranges','+NOUN'),('to','+PREP'),('avon','+guessed+NOUN'),('is','+VBPRES'),('that','+DET'),('right','+ADJ')],
- [('fill','+NOUN'),('it','+PRONPERS'),('with','+PREP'),('the','+DET'),('with','+PREP'),('the','+DET'),('oranges','+NOUN')],
- [('we','+PRONPERS'),('will','+VAUX'),('take','+VINF'),('we','+PRONPERS'),('will','+VAUX'),('take','+VINF'),('let','+VINF'),('us','+PRONPERS'),('say','+VPRES'),('one','+PRONONE'),('it','+PRONPERS'),('does','+VDPRES'),('not','+NOT'),('make','+VINF'),('a','+DET'),('difference','+NOUN'),('how','+WADV'),('fast','+ADV'),('a','+DET'),('train','+NOUN'),('goes','+VPRES'),('with','+PREP'),('one','+CARDONE'),('or','+COORD'),('two','+CARD')],
- [('let','+VPAST'),('us','+PRONPERS'),('just','+ADV'),('take','+VPRES'),('both','+QUANT'),('take','+VPRES'),('both','+QUANT'),('boxcars','+NOUN')],
- [('ok','+ADV'),('it','+PRONPERS'),('is','+VBPRES'),('seven','+CARD'),('a-m','+guessed+ADJ'),('fill','+NOUN'),('the','+DET'),('both','+QUANT'),('both','+COORD'),('up','+ADV'),('so','+ADV'),('that','+PRON'),('is','+VBPRES'),('nine','+CARD'),('a-

m','+guessed+NOUN'),('by','+PREP'),('the','+DET'),('time','+NOUN'),('they','+PRONPERS'),('are','+VBPRES'),('both','+QUANT'),('filled','+VPAST')],

- [('and','+COORD'),('then','+ADV'),('uh','+ITJ'),('I','+PRONPERS'),('want','+VPRES'),('to','+INFTO'),('send','+VINF'),('I','+PRONPERS'),('want','+VPRES'),('to','+INFTO'),('send','+VINF'),('uh','+ITJ'),('one','+CARDONE'),('boxcar','+NOUN'),('full','+ADV'),('of','+PREP'),('oranges','+NOUN'),('to','+PREP'),('bath','+NOUN')],
- [('right','+ADJ'),('right','+ADJ')],
- [('so','+ADV'),('so','+COSUB'),('they','+PRONPERS'),('will','+VAUX'),('be','+VBINF'),('arriving','+VPROG'),('at','+PREP'),('corning','+PARTPRES'),('at','+PREP'),('eight','+CARD'),('a-m','+guessed+ADJ')],
- [('so','+ADV'),('nine','+CARD'),('a-m','+guessed+ADJ'),('so','+COSUB'),('they','+PRONPERS'),('will','+VAUX'),('be','+VBINF'),('there','+ADV'),('um','+guessed+ADJ'),('at','+PREP'),('at','+PREP'),('eleven','+CARD'),('a-m','+guessed+ADJ'),('one','+PRONONE'),('will','+VAUX'),('arrive','+VINF'),('and','+COORD'),('the','+DET'),('other','+ADJPRON'),('one','+CARDONE'),('right','+ADJ'),('after','+PREP'),('that','+DET')],
- [('what','+WPRON'),('was','+VBPAST'),('the','+DET'),('the','+DET'),('goal','+NOUN'),('was','+VBPAST'),('to','+INFTO'),('get','+VINF')],
- [('ok','+ADV'),('so','+ADV'),('if','+COSUB'),('engine','+NOUN'),('if','+COSUB'),('engine','+NOUN'),('three','+CARD'),('or','+COORD'),('engine','+NOUN'),('two','+CARD'),('left','+ADJ'),('elmira','+guessed+ADJ')],
- [('um','+guessed+ADJ'),('to','+INFTO'),('make','+VINF'),('the','+DET'),('wait','+NOUN'),('the','+DET'),('the','+DET'),('only','+ADJ'),('thing','+NOUN'),('the','+DET'),('only','+ADJ'),('way','+NOUN'),('you','+PRONPERS'),('can','+VAUX'),('get','+VINF'),('orange','+ADJ'),('juice','+NOUN'),('is','+VBPRES'),('to','+INFTO'),('take','+VINF'),('oranges','+NOUN'),('to','+PREP'),('elmira','+guessed+ADJ'),('and','+COORD'),('make','+VINF'),('orange','+ADJ'),('juice','+NOUN'),('at','+PREP'),('the','+DET'),('orange','+ADJ'),('juice','+NOUN'),('factory','+NOUN')],
- [('you','+PRONPERS'),('mean','+VPRES'),('back','+ADV'),('to','+PREP'),('to','+PREP'),('avon','+guessed+ADJ')],
- [('because','+COSUB'),('uh','+ITJ'),('it','+PRONPERS'),('says','+VPRES'),('that','+COSUB'),('I','+PRONPERS'),('have','+VHPRES'),('to','+INFTO'),('get','+VINF'),('I','+PRONPERS'),('have','+VHPRES'),('to','+INFTO'),('get','+VINF'),('three','+CARD'),('boxcars','+NOUN'),('of','+PREP'),('bananas','+NOUN'),('to','+PREP'),('elmira','+guessed+ADJ'),('by','+PREP'),('nine','+CARD'),('o-clock','+guessed+ADJ'),('p-

m','+guessed+NOUN'),('but','+COSUB'),('I','+PRONPERS'),('can','+VAUX'),('only','+ADV'),('use','+VINF'),('engine','+NOUN'),('e','+PROP'),('two','+CARD')],

- [('and','+COORD'),('then','+ADV'),('and','+COORD'),('then','+ADV'),('uh','+ITJ'),('send','+VINF'),('it','+PRONPERS'),('all','+QUANT'),('back','+ADV'),('to','+PREP'),('dansville','+guessed+ADJ')],
- [('and','+COORD'),('then','+ADV'),('uh','+ITJ'),('and','+COORD'),('then','+ADV'),('send','+VINF'),('that','+COSUB'),('those','+DET'),('two','+CARD'),('tankers','+NOUN'),('to','+PREP'),('corning','+PARTPRES'),('then','+ADV'),('to','+PREP'),('dansville','+guessed+ADJ'),('then','+ADV'),('to','+PREP'),('avon','+guessed+ADJ')],
- [('to','+INFTO'),('have','+VHINF'),('um','+guessed+NOUN'),('two','+CARD'),('tankers','+NOUN'),('filled','+PARTPAST'),('with','+PREP'),('orange','+ADJ'),('juice','+NOUN')],
- [('ok','+ADJ'),('and','+COORD'),('then','+ADV'),('we','+PRONPERS'),('need','+VPRES'),('to','+PREP'),('we','+PRONPERS'),('are','+VBPRES'),('leaving','+VPROG'),('leaving','+VPROG'),('leaving','+VPROG'),('the','+DET'),('boxcars','+NOUN'),('and','+COORD'),('we','+PRONPERS'),('are','+VBPRES')],
- [('the','+DET'),('orange','+ADJ'),('juice','+NOUN'),('however','+ADV'),('our','+DET'),('problem','+NOUN'),('is','+VBPRES'),('that','+COSUB'),('all','+QUANT'),('the','+DET'),('other','+ADJPRON'),('engines','+NOUN'),('are','+VBPRES'),('in','+ADV'),('in','+PREP'),('undergoing','+PARTPRES'),('maintenance','+NOUN')],
- [('and','+COORD'),('since','+COSUB'),('there','+ADV'),('is','+VBPRES'),('a','+DET'),('time','+NOUN'),('limit','+NOUN'),('on','+PREP'),('the','+DET'),('bananas','+NOUN'),('I','+PRONPERS'),('think','+VPRES'),('we','+PRONPERS'),('should','+VAUX'),('work','+VINF'),('with','+PREP'),('that','+DET'),('problem','+NOUN'),('first','+ORD')],
- [('ok','+ADV'),('I','+PRONPERS'),('think','+VPRES'),('that','+COSUB'),('we','+PRONPERS'),('should','+VAUX'),('um','+guessed+ADJ'),('from','+PREP'),('elmira','+guessed+ADJ'),('take','+NOUN'),('engine','+NOUN'),('number','+NOUN'),('two','+CARD')],
- [('right','+ADJ'),('so','+COSUB'),('we','+PRONPERS'),('also','+ADV'),('need','+VPRES'),('three','+CARD'),('boxcars','+NOUN'),('empty','+ADJ'),('boxcars','+NOUN'),('to','+INFTO'),('bring','+VINF'),('with','+PREP'),('us','+PRONPERS')],
- [('yes','+NOUN'),('yes','+NOUN'),('and','+COORD'),('then','+ADV'),('as','+PREPADVAS'),('well','+ADV'),('um','+guessed+ADJ'),('um','+guessed+ADJ'),('dansville','+guessed+NOUN'),('is','+VBPRES'),('on','+PREP'),('the','+DET'),('faster','+ADJCMP'),('path','+NOUN'),('to','+PREP'),('avon','+guessed+ADJ'),('so','+COSUB'),('that','+DET'),('works','+NOUN'),('out','+ADV'),('well','+ADV')],
- [('so','+ADV'),('we','+PRONPERS'),('will','+VAUX'),('go','+VINF'),('so','+COSUB'),('we','+PRONPERS'),('will','+VAUX'),('take','+VINF'),('e','+PROP'),('two','+CARD'),('from','+PREP'),('elmira','+guessed+

ADJ'),('to','+PREP'),('corning','+PARTPRES'),('it','+PRONPERS'),('will','+VAUX'),('get','+VINF'),('there','+ADV'),('at','+PREP'),('two','+CARD'),('a-m','+guessed+NOUN'),('and','+COORD'),('then','+ADV'),('to','+PREP'),('dansville','+guessed+ADJ')],

- [('ok','+ADJ'),('so','+COSUB'),('it','+PRONPERS'),('will','+VAUX'),('be','+VBINF'),('seven','+CARD'),('a-m','+guessed+NOUN'),('by','+PREP'),('the','+DET'),('time','+NOUN'),('we','+PRONPERS'),('load','+VPRES'),('in','+PREP'),('load','+NOUN'),('the','+DET'),('bananas','+NOUN'),('then','+ADV'),('we','+PRONPERS'),('want','+VPRES'),('to','+INFTO'),('go','+VINF'),('back','+ADV'),('to','+PREP'),('elmira','+guessed+ADJ')],
- [('so','+ADV'),('at','+PREP'),('one','+CARDONE'),('p-m','+guessed+NOUN'),('here','+ADV'),('is','+VBPRES'),('the','+DET'),('we','+PRONPERS'),('will','+VAUX'),('have','+VHINF'),('our','+DET'),('three','+CARD'),('loaded','+ADJPAP'),('boxcars','+NOUN'),('of','+PREP'),('bananas','+NOUN'),('and','+COORD'),('we','+PRONPERS'),('will','+VAUX'),('have','+VHINF'),('two','+CARD'),('empty','+ADJ'),('tankers','+NOUN'),('of','+PREP'),('orange','+ADJ'),('juice','+NOUN')],
- [('and','+COORD'),('we','+PRONPERS'),('will','+VAUX'),('not','+NOT'),('but','+COSUB'),('we','+PRONPERS'),('will','+VAUX'),('not','+NOT'),('have','+VHINF'),('any','+QUANT'),('oranges','+NOUN'),('though','+ADV'),('to','+PREP'),('actually','+ADV'),('make','+VINF'),('the','+DET'),('orange','+ADJ'),('juice','+NOUN')],
- [('I','+PRONPERS'),('think','+VPRES'),('we','+PRONPERS'),('need','+VAUX'),('to','+INFTO'),('uh','+ITJ'),('I','+PRONPERS'),('need','+VAUX'),('we','+PRONPERS'),('need','+VPRES'),('to','+PREP'),('um','+guessed+ADJ'),('in','+PREP'),('the','+DET'),('initial','+ADJ'),('trip','+NOUN')],
- [('so','+ADV'),('when','+COSUB'),('we','+PRONPERS'),('come','+VINF'),('back','+ADV'),('we','+PRONPERS'),('can','+VAUX'),('have','+VHINF'),('those','+DET'),('so','+ADV'),('what','+WPRON'),('we','+PRONPERS'),('will','+VAUX'),('do','+VDINF'),('is','+VBPRES'),('is','+VBPRES'),('so','+ADV'),('we','+PRONPERS'),('drop','+VPRES'),('off','+PREP'),('the','+DET'),('two','+CARD'),('boxcars','+NOUN'),('in','+PREP'),('corning','+PARTPRES')],
- [('so','+ADV'),('we','+PRONPERS'),('are','+VBPRES'),('gonna','+guessed+NOUN'),('have','+VHPRES'),('to','+INFTO'),('leave','+VINF'),('leave','+VINF'),('leave','+VINF'),('those','+PRON'),('so','+COSUB'),('we','+PRONPERS'),('will','+VAUX'),('either','+COADV'),('have','+VHINF'),('to','+INFTO'),('leave','+VINF'),('leave','+VINF'),('leave','+VINF'),('the','+DET'),('bananas','+NOUN'),('at','+PREP'),('corning','+PARTPRES'),('or','+COORD'),('we','+PRONPERS'),('will','+VAUX'),('have','+VHINF'),('to','+INFTO'),('leave','+VINF'),('leave','+VINF'),('leave','+VINF'),('the','+DET'),('oranges','+NOUN'),('there','+ADV')],

- [('ok','+ADJ'),('so','+COSUB'),('we','+PRONPERS'),('will','+VAUX'),('get','+VINF'),('to','+PREP'),('so','+ADV'),('at','+PREP'),('corning','+PARTPRES'),('so','+ADV'),('we-d','+guessed+NOUN'),('be','+VBINF'),('at','+PREP'),('corning','+PARTPRES'),('at','+PREP'),('three','+CARD'),('a-m','+guessed+ADJ'),('we-d','+guessed+NOUN'),('be','+VBINF'),('in','+PREP'),('elmira','+guessed+ADJ'),('at','+PREP'),('five','+CARD'),('a-m','+guessed+ADJ'),('five','+CARD'),('p-m','+guessed+NOUN'),('I','+PRONPERS'),('I','+PRONPERS'),('mean','+VPRES')],
- [('ok','+ADV'),('so','+ADV'),('now','+ADV'),('to','+INFTO'),('get','+VINF'),('to','+PREP'),('avon','+guessed+ADJ'),('it','+PRONPERS'),('is','+VBPRES'),('gonna','+guessed+ADJ'),('take','+NOUN'),('two','+CARD'),('six','+CARD'),('more','+QUANTCMP'),('more','+QUANTCMP'),('hours','+NOUN'),('so','+COSUB'),('we','+PRONPERS'),('will','+VAUX'),('get','+VINF'),('to','+PREP'),('avon','+guessed+ADJ'),('at','+PREP'),('midnight','+NOUN'),('the','+DET'),('next','+ADJ'),('day','+NOUN')]
- [('um','+guessed+ADJ'),('so','+COSUB'),('I','+PRONPERS'),('guess','+VPRES'),('let','+VINF'),('us','+PRONPERS'),('see','+VPRES'),('bath','+NOUN'),('um','+guessed+ADJ'),('boxcars','+NOUN'),('are','+VBPRES'),('available','+ADJ'),('um','+guessed+NOUN'),('in','+PREP'),('dansville','+guessed+ADJ'),('and','+COORD'),('elmira','+guessed+ADJ'),('right','+ADJ')],
- [('and','+COORD'),('um','+guessed+ADJ'),('if','+COSUB'),('it','+PRONPERS'),('let','+VPRES'),('us','+PRONPERS'),('see','+VPRES'),('um','+guessed+NOUN'),('how','+WADV'),('long','+ADV'),('would','+VAUX'),('it','+PRONPERS'),('take','+VPRES'),('to','+INFTO'),('get','+VINF'),('an','+DET'),('engine','+NOUN'),('to','+INFTO'),('pick','+VINF'),('those','+DET'),('up','+ADV')],
- [('if','+COSUB'),('I','+PRONPERS'),('well','+ADV'),('let','+VPRES'),('us','+PRONPERS'),('see','+VPRES'),('if','+COSUB'),('two','+CARD'),('are','+VBPRES'),('in','+PREP'),('elmira','+guessed+ADJ'),('with','+PREP'),('the','+DET'),('engine','+NOUN'),('and','+COORD'),('I','+PRONPERS'),('would','+VAUX'),('have','+VHINF'),('to','+INFTO'),('go','+VINF'),('to','+PREP'),('danville','+guessed+NOUN'),('how','+WADV'),('long','+ADV'),('would','+VAUX'),('that','+ADV'),('take','+VPRES'),('to','+INFTO'),('get','+VINF'),('another','+DET'),('two','+CARD')],
- [('so','+ADV'),('you','+PRONPERS'),('want','+VPRES'),('to','+INFTO'),('take','+VINF'),('an','+DET'),('engine','+NOUN'),('and','+COORD'),('boxcar','+NOUN'),('from','+PREP'),('elmira','+guessed+ADJ'),('to','+PREP'),('dansville','+guessed+NOUN'),('is','+VBPRES'),('that','+COSUB'),('what','+WPRON'),('you','+PRONPERS'),('are','+VBPRES'),('asking','+VPROG')],
- [('um','+guessed+NOUN'),('I','+PRONPERS'),('guess','+VPRES'),('take','+VINF'),('one','+CARDONE'),('engine','+NOUN'),('and','+COORD'),('two','+CARD'),('boxcars','+NOUN')],

- [('um','+guessed+ADJ'),('yes','+NOUN'),('I','+PRONPERS'),('guess','+VPRES'),('um','+guessed+ADJ'),('or','+COORD'),('we','+PRONPERS'),('could','+VAUX'),('leave','+VINF'),('leave','+VINF'),('leave','+VINF'),('now','+ADV'),('I','+PRONPERS'),('I','+PRONPERS'),('guess','+VPRES'),('it','+PRONPERS'),('does','+VDPRES'),('not','+NOT'),('matter','+VINF'),('um','+guessed+NOUN'),('does','+VDPRES'),('that','+DET'),('sound','+NOUN'),('like','+PREP'),('it','+PRONPERS'),('makes','+VPRES'),('sense','+NOUN'),('to','+INFTO'),('transport','+VINF')],
- [('no','+ADV'),('I','+PRONPERS'),('just','+ADV'),('it','+PRONPERS'),('just','+ADV'),('makes','+VPRES'),('it','+PRONPERS'),('clearer','+ADJCMP')],
- [('so','+ADV'),('yes','+VINF'),('um','+guessed+ADJ'),('I','+PRONPERS'),('guess','+VPRES'),('the','+DET'),('instructions','+NOUN'),('are','+VBPRES'),('would','+VAUX'),('I','+PRONPERS'),('be','+VBINF'),('giving','+VPROG'),('you','+PRONPERS'),('instructions','+NOUN'),('I','+PRONPERS'),('do','+VDPRES'),('not','+NOT'),('even','+ADV'),('know','+VINF')]
- [('um','+guessed+ADJ'),('two','+CARD'),('since','+COSUB'),('we','+PRONPERS'),('there','+ADV'),('is','+VBPRES'),('only','+ADV'),('two','+CARD'),('available','+ADJ'),('in','+PREP'),('elmira','+guessed+ADJ')],
- [('ok','+ADV'),('you','+PRONPERS'),('can','+VAUX'),('only','+ADV'),('carry','+VINF'),('um','+guessed+ADJ'),('you','+PRONPERS'),('can','+VAUX'),('carry','+VINF'),('three','+CARD'),('empty','+ADJ'),('boxcars','+NOUN'),('that','+PRONREL'),('would','+VAUX'),('be','+VBINF'),('fine','+ADJ'),('but','+COSUB'),('once','+ADV'),('you','+PRONPERS'),('get','+VPRES'),('them','+PRONPERS'),('loaded','+VPAST'),('you','+PRONPERS'),('can','+VAUX'),('carry','+VINF'),('as','+PREPADVAS'),('many','+QUANT'),('um','+guessed+ADJ'),('unloaded','+ADJPAP'),('boxcars','+NOUN'),('as','+PREPADVAS'),('you','+PRONPERS'),('want','+VPRES'),('but','+COSUB'),('once','+ADV'),('you','+PRONPERS'),('get','+VPRES'),('them','+PRONPERS'),('loaded','+VPAST'),('you','+PRONPERS'),('can','+VAUX'),('only','+ADV'),('carry','+VINF'),('three','+CARD'),('at','+PREP'),('a','+DET'),('time','+NOUN'),('um','+guessed+NOUN'),('you','+PRONPERS'),('can','+VAUX'),('carry','+VINF'),('three','+CARD'),('empty','+ADJ'),('boxcars','+NOUN'),('that','+PRONREL'),('would','+VAUX'),('be','+VBINF'),('fine','+ADJ'),('but','+COSUB'),('once','+ADV'),('you','+PRONPERS'),('get','+VPRES'),('them','+PRONPERS'),('loaded','+VPAST'),('you','+PRONPERS'),('can','+VAUX'),('carry','+VINF'),('as','+PREPADVAS'),('many','+QUANT'),('um','+guessed+ADJ'),('unloaded','+ADJPAP'),('boxcars','+NOUN'),('as','+PREPADVAS'),('you','+PRONPERS'),('want','+VPRES'),('but','+COSUB'),('once','+ADV'),('you','+PRONPERS'),('get','+VPRES'),('them','+PRONPERS'),('loaded','+VPAST'),('you','+PRONPERS'),('can','+VAUX'),('only','+ADV'),('carry','+VINF'),('three','+CARD'),('at','+PREP'),('a','+DET'),('time','+NOUN')]

- [('pick','+NOUN'),('up','+ADV'),('and','+COORD'),('load','+NOUN'),('two','+CARD'),('um','+guessed+ADJ'),('the','+DET'),('two','+CARD'),('uh','+ITJ'),('boxcars','+NOUN'),('on','+PREP'),('engine','+NOUN'),('two','+CARD'),('with','+PREP'),('oranges','+NOUN'),('and','+COORD'),('then','+ADV'),('continue','+VINF'),('to','+PREP'),('bath','+NOUN')],
- [('um','+guessed+NOUN'),('at','+PREP'),('the','+DET'),('same','+ADJPRON'),('time','+NOUN'),('engine','+NOUN'),('three','+CARD'),('would','+VAUX'),('go','+VINF'),('um','+guessed+ADJ'),('from','+PREP'),('elmira','+guessed+ADJ'),('to','+PREP'),('corning','+PARTPRES'),('to','+PREP'),('dansville','+guessed+ADJ')],
- [('so','+ADV'),('it','+PRONPERS'),('takes','+VPRES'),('ooh','+ITJ'),('if','+COSUB'),('you','+PRONPERS'),('leave','+VPRES'),('leave','+VPRES'),('leave','+VPRES'),('elmira','+guessed+NOUN'),('at','+PREP'),('twelve','+CARD'),('a-m','+guessed+NOUN'),('with','+PREP'),('the','+DET'),('two','+CARD'),('boxcars','+NOUN'),('you','+PRONPERS'),('get','+VPRES'),('to','+PREP'),('corning','+PARTPRES'),('at','+PREP'),('two','+CARD'),('a-m','+guessed+ADJ')],
- [('um','+guessed+NOUN'),('and','+COORD'),('the','+DET'),('time','+NOUN'),('there','+ADV'),('is','+VBPRES'),('no','+ADV'),('if','+COSUB'),('we','+PRONPERS'),('left','+VPAST'),('now','+ADV'),('the','+DET'),('time','+NOUN'),('wouldn-t','+guessed+NOUN'),('affect','+VPRES'),('that','+PRON'),('would','+VAUX'),('it','+PRONPERS')],
- [('five','+CARD'),('a-m','+guessed+ADJ'),('ok','+NOUN'),('is','+VBPRES'),('it','+PRONPERS'),('faster','+ADVCMP'),('for','+PREP'),('those','+PRON'),('for','+PREP'),('that','+DET'),('engine','+NOUN'),('to','+INFTO'),('drop','+VINF'),('off','+PREP'),('those','+DET'),('two','+CARD'),('those','+DET'),('two','+CARD'),('boxcars','+NOUN'),('travel','+NOUN'),('back','+ADV'),('to','+PREP'),('dansville','+guessed+ADJ'),('than','+COTHAN'),('um','+guessed+ADJ'),('to','+INFTO'),('have','+VHINF'),('engine','+NOUN'),('three','+CARD')],
- [('ok','+ADV'),('so','+ADV'),('then','+ADV'),('yes','+VINF'),('I','+PRONPERS'),('guess','+VPRES'),('just','+ADV'),('have','+VHINF'),('the','+DET'),('the','+DET'),('original','+NOUN'),('take','+VPRES'),('the','+DET'),('two','+CARD'),('have','+VHINF'),('them','+PRONPERS'),('leave','+VPRES'),('leave','+VPRES'),('leave','+VPRES'),('at','+PREP'),('the','+DET'),('same','+ADJPRON'),('time','+NOUN')],
- [('if','+COSUB'),('um','+guessed+ADJ'),('oh','+NOUN'),('and','+COORD'),('ok','+ADJ'),('uh','+ITJ'),('and','+COORD'),('it','+PRONPERS'),('is','+VBPRES'),('still','+ADV'),('the','+DET'),('case','+NOUN'),('that','+COSUB'),('um','+guessed+ADJ'),('only','+ADV'),('four','+CARD'),('boxcars','+NOUN'),('with','+PREP'),('cargo','+NOUN'),('can','+VAUX'),('go','+VINF'),('at','+PREP'),('a','+DET'),('time','+NOUN'),('is','+VBPRES'),('that','+DET'),('right','+ADJ')],

- [('four','+CARD'),('hours','+NOUN'),('and','+COORD'),('then','+ADV'),('from','+PREP'),('oh','+NOUN'),('ok','+ADV'),('let','+VPRES'),('us','+PRONPERS'),('see','+VPRES'),('and','+COORD'),('how','+WADV'),('about','+ADV'),('avon','+guessed+ADJ'),('to','+PREP'),('dansville','+guessed+ADJ')],
- [('go','+NOUN'),('back','+ADV'),('hm','+guessed+NOUN'),('how','+WADV'),('about','+ADV'),('um','+guessed+NOUN'),('how','+WADV'),('long','+ADV'),('is','+VBPRES'),('it','+PRONPERS'),('from','+PREP'),('elmira','+guessed+ADJ'),('to','+PREP'),('dansville','+guessed+ADJ')],
- [('so','+ADV'),('why','+WADV'),('do','+VDPRES'),('not','+NOT'),('uh','+ITJ'),('I','+PRONPERS'),('send','+VPRES'),('engine','+NOUN'),('two','+CARD'),('with','+PREP'),('two','+CARD'),('boxcars','+NOUN'),('to','+PREP'),('corning','+PARTPRES')],
- [('so','+ADV'),('that','+COSUB'),('will','+VAUX'),('take','+VINF'),('uh','+ITJ'),('but','+COSUB'),('no','+DET'),('four','+CARD'),('hours','+NOUN'),('is','+VBPRES'),('that','+DET'),('right','+ADJ')],
- [('ok','+ADV'),('I','+PRONPERS'),('guess','+VINF'),('how','+WADV'),('it','+PRONPERS'),('takes','+VPRES'),('how','+WADV'),('long','+ADJ'),('to','+INFTO'),('get','+VINF'),('from','+PREP'),('elmira','+guessed+ADJ'),('to','+PREP'),('corning','+PARTPRES')],
- [('ok','+ADV'),('I','+PRONPERS'),('guess','+VPRES'),('um','+guessed+ADJ'),('well','+ADV'),('how','+WADV'),('many','+QUANT'),('it','+PRONPERS'),('is','+VBPRES'),('how','+WADV'),('long','+ADV'),('is','+VBPRES'),('from','+PREP'),('corning','+PARTPRES'),('to','+PREP'),('elmira','+guessed+ADJ'),('an','+DET'),('hour','+NOUN')],
- [('um','+guessed+ADJ'),('engine','+NOUN'),('three','+CARD'),('will','+VAUX'),('be','+VBINF'),('taking','+VPROG'),('three','+CARD'),('um','+guessed+NOUN'),('no','+ADV'),('two','+CARD'),('boxcars','+NOUN'),('to','+PREP'),('elmira','+guessed+ADJ'),('which-ll','+guessed+ADJ'),('take','+NOUN'),('is','+VBPRES'),('that','+COSUB'),('two','+CARD'),('hours','+NOUN')],
- [('ok','+ADV'),('so','+ADV'),('then','+ADV'),('in','+PREP'),('so','+ADV'),('by','+PREP'),('the','+DET'),('time','+NOUN'),('um','+guessed+NOUN'),('by','+PREP'),('the','+DET'),('time','+NOUN'),('engine','+NOUN'),('three','+CARD'),('arrives','+VPRES'),('at','+PREP'),('elmira','+guessed+ADJ')],
- [('um','+guessed+ADJ'),('alright','+guessed+NOUN'),('so','+ADV'),('engine','+NOUN'),('three','+CARD'),('will','+NOUN'),('engine','+NOUN'),('three','+CARD'),('will','+VAUX'),('stay','+VINF'),('in','+PREP'),('corning','+PARTPRES'),('and','+COORD'),('load','+VINF'),('the','+DET'),('oranges','+NOUN'),('and','+COORD'),('it','+PRONPERS'),('will','+VAUX'),('have','+VHINF'),('two','+CARD'),('boxcars','+NOUN'),('of','+PREP'),('oranges','+NOUN'),('loaded','+PARTPAST'),('at','+PREP'),('three','+CARD'),('a-m','+guessed+ADJ')],

- [('um','+guessed+NOUN'),('it','+PRONPERS'),('will','+VAUX'),('be','+VBINF'),('there','+ADV'),('it','+PRONPERS'),('will','+VAUX'),('get','+VINF'),('to','+PREP'),('dansville','+guessed+ADJ'),('at','+PREP'),('three','+CARD'),('a-m','+guessed+NOUN'),('and','+COORD'),('then','+ADV'),('you','+PRONPERS'),('I','+PRONPERS'),('want','+VPRES'),('do','+VDPRES'),('you','+PRONPERS'),('take','+VPRES'),('want','+NOUN'),('to','+INFTO'),('take','+VINF'),('those','+DET'),('back','+NOUN'),('to','+PREP'),('elmira','+guessed+ADJ')],
- [('did','+VDPAST'),('you','+PRONPERS'),('the','+DET'),('three','+CARD'),('boxcars','+NOUN'),('from','+PREP'),('dansville','+guessed+NOUN'),('do','+VDPRES'),('you','+PRONPERS'),('want','+VPRES'),('them','+PRONPERS'),('to','+INFTO'),('stop','+VINF'),('in','+PREP'),('corning','+PARTPRES'),('and','+COORD'),('load','+NOUN'),('and','+COORD'),('then','+ADV'),('go','+VPRES'),('to','+PREP'),('elmira','+guessed+ADJ')],
- [('ok','+ADJ'),('so','+COSUB'),('those','+DET'),('three','+CARD'),('will','+VAUX'),('be','+VBINF'),('back','+ADV'),('in','+PREP'),('elmira','+guessed+ADJ'),('at','+PREP'),('seven','+CARD'),('a-m','+guessed+ADJ')],
- [('ok','+ADV'),('can','+VAUX'),('I','+PRONPERS'),('run','+VINF'),('tracks','+NOUN'),('I','+PRONPERS'),('can','+VAUX'),('run','+VINF'),('trains','+NOUN'),('on','+PREP'),('the','+DET'),('in','+ADV'),('the','+DET'),('opposite','+ADJ'),('direction','+NOUN'),('right','+ADJ')],
- [('so','+ADV'),('if','+COSUB'),('they','+PRONPERS'),('both','+QUANT'),('ok','+ADJ'),('both','+COORD'),('engine','+NOUN'),('two','+CARD'),('and','+COORD'),('engine','+NOUN'),('three','+CARD'),('go','+NOUN'),('from','+PREP'),('elmira','+guessed+ADJ'),('to','+PREP'),('corning','+PARTPRES'),('they','+PRONPERS'),('get','+VPRES'),('there','+ADV'),('by','+PREP'),('two','+CARD'),('a-m','+guessed+ADJ')],
- [('I','+PRONPERS'),('think','+VPRES'),('I','+PRONPERS'),('am','+VBPRES'),('gonna','+guessed+ADJ'),('say','+NOUN'),('this','+PRON'),('is','+VBPRES'),('impossible','+ADJ'),('unless','+COSUB'),('how','+WADV'),('long','+ADV'),('does','+VDPRES'),('it','+PRONPERS'),('take','+VPRES'),('from','+PREP'),('get','+NOUN'),('to','+PREP'),('corning','+PARTPRES'),('to','+PREP'),('bath','+NOUN')],
- [('and','+COORD'),('then','+ADV'),('um','+guessed+ADJ'),('so','+COSUB'),('that','+PRON'),('could','+VAUX'),('get','+VINF'),('there','+ADV'),('by','+PREP'),('nine','+CARD'),('ten','+CARD'),('no','+ADV'),('by','+PREP'),('twelve','+CARD'),('three','+CARD')],

- [('ok','+ADJ'),('and','+COORD'),('then','+ADV'),('get','+NOUN'),('to','+PREP'),('corning','+PARTPRES'),('that','+COSUB'),('will','+VAUX'),('take','+VINF'),('it','+PRONPERS'),('will','+VAUX'),('be','+VBINF'),('there','+ADV'),('at','+PREP'),('six','+CARD'),('a-m','+guessed+NOUN'),('it','+PRONPERS'),('can','+VAUX'),('load','+VINF'),('the','+DET'),('oranges','+

NOUN'),('and','+COORD'),('be','+VBINF'),('in','+PREP'),('elmira','+guessed+ADJ'),('at','+PREP'),('nine','+CARD'),('a-m','+guessed+ADJ')],

- [('and','+COORD'),('I','+PRONPERS'),('need','+VAUX'),('let','+VINF'),('us','+PRONPERS'),('see','+VPRES'),('engine','+NOUN'),('one','+CARDONE'),('um','+guessed+ADJ'),('to','+INFTO'),('leave','+VINF'),('leave','+VINF'),('leave','+VINF'),('let','+VINF'),('us','+PRONPERS'),('see','+VPRES'),('ok','+ADJ'),('um','+guessed+NOUN'),('I','+PRONPERS'),('need','+VPRES'),('two','+CARD'),('boxcars','+NOUN'),('to','+INFTO'),('go','+VINF'),('to','+PREP'),('avon','+guessed+ADJ'),('from','+PREP'),('elmira','+guessed+ADJ'),('un','+guessed+ADJ'),('contre','+guessed+ADJ'),('exemple','+guessed+ADJ')],
- [('um','+guessed+ADJ'),('yes','+NOUN'),('I','+PRONPERS'),('would','+VAUX'),('like','+VINF'),('one','+CARDONE'),('tanker','+NOUN'),('um','+guessed+ADJ'),('one','+CARDONE'),('engine','+NOUN'),('to','+INFTO'),('leave','+VINF'),('leave','+VINF'),('leave','+VINF'),('elmira','+guessed+ADJ'),('with','+PREP'),('one','+CARDONE'),('tank','+NOUN'),('of','+PREP'),('orange','+ADJ'),('juice','+NOUN'),('to','+INFTO'),('go','+VINF'),('to','+PREP'),('avon','+guessed+ADJ'),('by','+PREP'),('three','+CARD'),('p-m','+guessed+ADJ')],
- [('oh','+NOUN'),('ok','+ADV'),('it','+PRONPERS'),('depends','+VPRES'),('on','+ADV'),('to','+INFTO'),('get','+VINF'),('a','+DET'),('tank','+NOUN'),('of','+PREP'),('orange','+ADJ'),('juice','+NOUN'),('first','+ORD'),('you','+PRONPERS'),('need','+VAUX'),('to','+INFTO'),('pick','+VINF'),('up','+ADV'),('some','+QUANT'),('oranges','+NOUN'),('take','+VPRES'),('them','+PRONPERS'),('to','+PREP'),('elmira','+guessed+ADJ')],
- [('turn','+NOUN'),('it','+PRONPERS'),('into','+PREP'),('orange','+ADJ'),('juice','+NOUN')],
- [('we','+PRONPERS'),('need','+VPRES'),('one','+CARDONE'),('engine','+NOUN'),('to','+INFTO'),('go','+VINF'),('to','+PREP'),('elmira','+guessed+ADJ'),('to','+PREP'),('corning','+PARTPRES'),('um','+guessed+ADJ'),('leave','+NOUN'),('at','+PREP'),('midnight','+NOUN')],
- [('and','+COORD'),('pick','+NOUN'),('up','+PREP'),('um','+guessed+ADJ'),('the','+DET'),('I','+PRONPERS'),('guess','+VPRES'),('the','+DET'),('entire','+ADJ'),('um','+guessed+ADJ'),('pick','+NOUN'),('up','+PREP'),('the','+DET'),('load','+NOUN'),('of','+PREP'),('oranges','+NOUN'),('at','+PREP'),('corning','+PARTPRES'),('and','+COORD'),('bring','+VPRES'),('them','+PRONPERS'),('back','+ADV'),('to','+PREP'),('elmira','+guessed+ADJ')],
- [('ok','+ADJ'),('um','+guessed+NOUN'),('and','+COORD'),('how','+WADV'),('long','+ADV'),('does','+VDPRES'),('it','+PRONPERS'),('can','+VAUX'),('how','+WADV'),('long','+ADV'),('does','+VDPRES'),('it','+PRONPERS'),('take','+VPRES'),('to','+INFTO'),('convert','+VINF'),('the','+DET'),('oranges','+NOUN'),('into','+PREP'),('orange','+ADJ'),('juice','+NOUN')],

- [('um','+guessed+NOUN'),('it','+PRONPERS'),('could','+VAUX'),('you','+PRONPERS'),('it','+PRONPERS'),('can','+VAUX'),('take','+VINF'),('you','+PRONPERS'),('could','+VAUX'),('take','+VINF'),('the','+DET'),('eight','+CARD'),('hour','+NOUN'),('way','+NOUN'),('or','+COORD'),('the','+DET'),('six','+CARD'),('hour','+NOUN'),('way','+NOUN'),('I','+PRONPERS'),('assume','+VPRES'),('you','+PRONPERS'),('want','+VPRES'),('to','+INFTO'),('take','+VINF'),('the','+DET'),('six','+CARD'),('hour','+NOUN'),('way','+NOUN'),('right','+ADJ')],
- [('so','+ADV'),('you','+PRONPERS'),('are','+VBPRES'),('trying','+VPROG'),('to','+INFTO'),('get','+VINF'),('what','+WPRON'),('you','+PRONPERS'),('need','+VPRES'),('is','+VBPRES'),('to','+INFTO'),('get','+VINF'),('one','+CARDONE'),('engine','+NOUN'),('of','+PREP'),('orange','+ADJ'),('juice','+NOUN'),('from','+PREP'),('elmira','+guessed+ADJ'),('to','+PREP'),('avon','+guessed+ADJ'),('by','+PREP'),('three','+CARD'),('p-m','+guessed+ADJ')],
- [('ok','+ADJ'),('and','+COORD'),('what','+WDET'),('time','+NOUN'),('would','+VAUX'),('that','+ADV'),('er','+guessed+ADJ'),('what','+WDET'),('time','+NOUN'),('would','+VAUX'),('the','+DET'),('boxcar','+NOUN'),('arrive','+VPRES'),('in','+PREP'),('avon','+guessed+ADJ')],
- [('um','+guessed+NOUN'),('it','+PRONPERS'),('could','+VAUX'),('leave','+VINF'),('leave','+VINF'),('leave','+VINF'),('there','+ADV'),('at','+PREP'),('noon','+NOUN'),('after','+PREP'),('the','+DET'),('orange','+ADJ'),('juice','+NOUN'),('had','+VHPAST'),('gotten','+VPAP'),('there','+ADV'),('and','+COORD'),('get','+NOUN'),('to','+PREP'),('but','+COSUB'),('then','+ADV'),('it','+PRONPERS'),('wouldn-t','+guessed+ADJ'),('get','+NOUN'),('to','+PREP'),('corning','+PARTPRES'),('by','+ADV'),('until','+PREP'),('four','+CARD'),('p-m','+guessed+ADJ')],
- [('um','+guessed+ADJ'),('yes','+NOUN'),('I','+PRONPERS'),('need','+VAUX'),('to','+INFTO'),('send','+VINF'),('let','+VINF'),('us','+PRONPERS'),('see','+VINF'),('how','+WADV'),('many','+QUANT'),('um','+guessed+ADJ'),('boxcars','+NOUN'),('can','+VPRES'),('one','+CARDONE'),('engine','+NOUN'),('take','+VPRES')],
- [('um','+guessed+ADJ'),('if','+COSUB'),('the','+DET'),('boxcars','+NOUN'),('are','+VBPRES'),('unloaded','+VPAP'),('as','+PREPADVAS'),('many','+QUANT'),('as','+PREPADVAS'),('it','+PRONPERS'),('as','+PREPADVAS'),('it','+PRONPERS'),('can','+VAUX'),('and','+COORD'),('if','+COSUB'),('they','+PRONPERS'),('are','+VBPRES'),('loaded','+VPAP'),('it','+PRONPERS'),('can','+VAUX'),('carry','+VINF'),('three','+CARD')],
- [('ok','+NOUN'),('which','+PRONREL'),('that','+PRON'),('is','+VBPRES'),('the','+DET'),('that','+ADV'),('is','+VBPRES'),('your','+DET'),('ultimate','+ADJ'),('goal','+NOUN')],
- [('um','+guessed+NOUN'),('so','+ADV'),('if','+COSUB'),('we','+PRONPERS'),('if','+COSUB'),('we','+PRONPERS'),('leave','+VPRES'),('leave','+VPRES'),('leave','+VPRES'),('um','+guessed+ADJ'),('if','+COS

UB'),('the','+DET'),('boxcars','+NOUN'),('leave','+NOUN'),('elmira','+guessed+NOUN'),('by','+PREP'),('um','+guessed+ADJ'),('at','+PREP'),('midnight','+NOUN')],

- [('now','+ADV'),('originally','+ADV'),('you','+PRONPERS'),('said','+VPAST'),('that','+COSUB'),('the','+DET'),('engine','+NOUN'),('can','+VAUX'),('take','+VINF'),('as','+PREPADVAS'),('many','+QUANT'),('um','+guessed+NOUN'),('it','+PRONPERS'),('can','+VAUX'),('take','+VINF'),('up','+ADV'),('to','+PREP'),('three','+CARD'),('loaded','+ADJPAP'),('boxcars','+NOUN')],
- [('ok','+ADJ'),('ok','+ADJ'),('so','+COSUB'),('we','+PRONPERS'),('send','+VPRES'),('the','+DET'),('two','+CARD'),('boxcars','+NOUN'),('from','+PREP'),('elmira','+guessed+ADJ'),('to','+PREP'),('avon','+guessed+ADJ'),('and','+COORD'),('pick','+NOUN'),('up','+PREP'),('the','+DET'),('two','+CARD'),('um','+guessed+NOUN'),('let','+VPRES'),('us','+PRONPERS'),('see','+VPRES'),('we','+PRONPERS'),('pick','+VPRES'),('up','+PREP'),('the','+DET'),('two','+CARD'),('loads','+NOUN'),('of','+PREP'),('bananas','+NOUN'),('and','+COORD'),('bring','+VPRES'),('them','+PRONPERS'),('to','+PREP'),('dansville','+guessed+ADJ')],
- [('ok','+ADJ'),('um','+guessed+NOUN'),('no','+ADV'),('let','+VPRES'),('us','+PRONPERS'),('see','+VPRES')]
- [('we','+PRONPERS'),('need','+VPRES'),('to','+PREP'),('we','+PRONPERS'),('should','+VAUX'),('have','+VHINF'),('both','+QUANT'),('of','+PREP'),('the','+DET'),('boxcars','+NOUN'),('at','+PREP'),('dansville','+guessed+ADJ'),('by','+PREP'),('noon','+NOUN'),('so','+ADV')],
- [('we','+PRONPERS'),('need','+VPRES'),('a','+DET'),('shorter','+ADJCMP'),('route','+NOUN'),('from','+PREP'),('we','+PRONPERS'),('need','+VPRES'),('to','+PREP'),('um','+guessed+NOUN'),('manage','+VPRES'),('to','+INFTO'),('get','+VINF'),('the','+DET'),('bananas','+NOUN'),('to','+PREP'),('dansville','+guessed+NOUN'),('more','+QUANTCMP'),('more','+QUANTCMP'),('quickly','+ADV'),('um','+guessed+ADJ')],
- [('what','+WPRON'),('you','+PRONPERS'),('could','+VAUX'),('do','+VDINF'),('is','+VBPRES'),('you','+PRONPERS'),('could','+VAUX'),('take','+VINF'),('the','+DET'),('boxcars','+NOUN'),('from','+PREP'),('elmira','+guessed+ADJ'),('load','+NOUN'),('the','+DET'),('oranges','+NOUN'),('at','+PREP'),('corning','+PARTPRES'),('and','+COORD'),('drop','+NOUN'),('off','+PREP'),('the','+DET'),('oranges','+NOUN'),('at','+PREP'),('dansville','+guessed+ADJ'),('on','+PREP'),('your','+DET'),('way','+NOUN'),('to','+PREP'),('avon','+guessed+ADJ')],
- [('and','+COORD'),('how','+WADV'),('many','+QUANT'),('boxcars','+NOUN'),('of','+PREP'),('do','+VDPRES'),('you','+PRONPERS'),('need','+VAUX'),('to','+INFTO'),('get','+VINF'),('to','+PREP'),('avon','+guessed+ADJ'),('to','+INFTO'),('load','+VINF'),('bananas','+NOUN')],

- [('ok','+ADJ'),('and','+COORD'),('um','+guessed+ADJ'),('so','+COSUB'),('they','+PRONPERS'),('will','+VAUX'),('be','+VBINF'),('unloaded','+VPAP'),('by','+PREP'),('noon','+NOUN')],
- [('and','+COORD'),('get','+VINF'),('the','+DET'),('that','+ADV'),('one','+CARDONE'),('boxcar','+NOUN'),('of','+PREP'),('oranges','+NOUN'),('and','+COORD'),('get','+VPRES'),('it','+PRONPERS'),('to','+PREP'),('dansville','+guessed+ADJ')],
- [('I','+PRONPERS'),('you','+PRONPERS'),('know','+VPRES'),('I','+PRONPERS'),('will','+VAUX'),('use','+VINF'),('I','+PRONPERS'),('will','+VAUX'),('do','+VDINF'),('that','+PRON'),('drop','+VPRES'),('them','+PRONPERS'),('off','+ADV'),('and','+COORD'),('I','+PRONPERS'),('will','+VAUX'),('have','+VHINF'),('that','+DET'),('one','+CARDONE'),('engine','+NOUN'),('available','+ADJ')],
- [('and','+COORD'),('pick','+NOUN'),('up','+PREP'),('two','+CARD'),('unloaded','+VPAST'),('two','+CARD'),('empty','+ADJ'),('boxcars','+NOUN')],
- [('yes','+NOUN'),('you','+PRONPERS'),('can','+VPRES'),('um','+guessed+ADJ'),('I','+PRONPERS'),('have','+VHPRES'),('to','+INFTO'),('pick','+VINF'),('up','+ADV'),('two','+CARD'),('boxcars','+NOUN'),('of','+PREP'),('oranges','+NOUN'),('I','+PRONPERS'),('have','+VHPRES'),('to','+INFTO'),('make','+VINF'),('it','+PRONPERS'),('into','+PREP'),('I','+PRONPERS'),('can','+VAUX'),('make','+VINF'),('into','+PREP'),('two','+CARD'),('tankers','+NOUN'),('of','+PREP'),('orange','+ADJ'),('juice','+NOUN')],
- [('so','+ADV'),('we','+PRONPERS'),('have','+VHPRES'),('eleven','+CARD'),('hours','+NOUN'),('wait','+NOUN'),('is','+VBPRES'),('that','+COSUB'),('right','+ADV'),('no','+ADV'),('we','+PRONPERS'),('have','+VHPRES'),('thirteen','+CARD'),('hours','+NOUN')],
- [('well','+ADV'),('you','+PRONPERS'),('can','+VAUX'),('carry','+VINF'),('it','+PRONPERS'),('is','+VBPRES'),('loaded','+VPAP'),('boxcars','+NOUN'),('you','+PRONPERS'),('can','+VAUX'),('pull','+VINF'),('as','+PREPADVAS'),('many','+QUANT'),('uh','+ITJ'),('um','+guessed+ADJ'),('empty','+ADJ'),('things','+NOUN'),('as','+PREPADVAS'),('you','+PRONPERS'),('as','+PREPADVAS'),('you','+PRONPERS'),('want','+VPRES')],
- [('oh','+ITJ'),('wait','+VINF'),('a','+DET'),('minute','+NOUN'),('you','+PRONPERS'),('you','+PRONPERS'),('have','+VHPRES'),('to','+INFTO')],
- [('how','+WADV'),('long','+ADV'),('will','+VAUX'),('it','+PRONPERS'),('take','+VPRES'),('engine','+NOUN'),('one','+PRONONE'),('to','+PREP'),('get','+NOUN'),('to','+PREP'),('danville','+guessed+ADJ'),('dansville','+guessed+ADJ')],
- [('ok','+ADJ'),('and','+COORD'),('how','+WADV'),('long','+ADJ'),('will','+NOUN'),('that','+PRONREL'),('take','+VPRES'),('will','+VAUX'),('it','+PRONPERS'),('take','+VPRES'),('for','+PREP'),('um','+guessed+ADJ'),('engine','+NOUN'),('one','+PRONONE'),('at','+PREP'),('dansville','+guessed+ADJ')],

- [('so','+ADV'),('four','+CARD'),('hours','+NOUN'),('in','+PREP'),('all','+QUANT'),('from','+PREP'),('starting','+PARTPRES'),('from','+PREP'),('avon','+guessed+ADJ')],
- [('how','+WADV'),('long','+ADV'),('will','+VAUX'),('it','+PRONPERS'),('take','+VPRES'),('me','+PRONPERS'),('to','+INFTO'),('uh','+ITJ'),('carry','+VINF'),('use','+NOUN'),('engine','+NOUN'),('two','+CARD'),('to','+PREP'),('carry','+NOUN'),('three','+CARD'),('box','+NOUN'),('cars','+NOUN'),('of','+PREP'),('loaded','+ADJPAP'),('oranges','+NOUN'),('from','+PREP'),('corning','+PARTPRES'),('to','+PREP'),('bath','+NOUN')],
- [('and','+COORD'),('bring','+VPRES'),('that','+PRON'),('to','+PREP'),('corning','+PARTPRES'),('and','+COORD'),('load','+VPRES'),('uh','+ITJ'),('more','+QUANTCMP'),('more','+QUANTCMP'),('oranges','+NOUN'),('two','+CARD'),('more','+QUANTCMP'),('more','+QUANTCMP'),('boxcars','+NOUN'),('of','+PREP'),('oranges','+NOUN')],
- [('um','+guessed+NOUN'),('I','+PRONPERS'),('think','+VPRES'),('maybe','+ADV'),('using','+ADJING'),('engine','+NOUN'),('e','+PROP'),('one','+CARDONE'),('and','+COORD'),('taking','+PARTPRES'),('the','+DET'),('take','+NOUN'),('making','+PARTPRES'),('the','+DET'),('trip','+NOUN'),('to','+PREP'),('um','+guessed+ADJ'),('dansville','+guessed+ADJ'),('multiple','+ADJ')],
- [('and','+COORD'),('then','+ADV'),('that','+COSUB'),('takes','+VPRES'),('what','+WDET'),('three','+CARD'),('hours','+NOUN')],
- [('alright','+guessed+NOUN'),('we','+PRONPERS'),('I','+PRONPERS'),('want','+VPRES'),('be','+VBINF'),('loading','+VPROG'),('two','+CARD'),('at','+PREP'),('the','+DET'),('banana','+NOUN'),('warehouse','+NOUN'),('but','+COSUB')],
- [('ok','+ADJ'),('oh','+NOUN'),('this','+ADV'),('this','+PRON'),('I','+PRONPERS'),('do','+VDPRES'),('not','+NOT'),('know','+VINF'),('if','+COSUB'),('this','+PRON'),('is','+VBPRES'),('going','+VPROG'),('work','+NOUN'),('now','+ADV')],
- [('so','+ADV'),('we','+PRONPERS'),('could','+VAUX'),('just','+ADV'),('we','+PRONPERS'),('do','+VDPRES'),('not','+NOT'),('have','+VHINF'),('to','+INFTO'),('wait','+VINF'),('for','+PREP'),('them','+PRONPERS'),('to','+INFTO'),('be','+VBINF'),('unloaded','+VPAP'),('we','+PRONPERS'),('just','+ADV'),('unhitch','+VINF'),('them','+PRONPERS')],
- [('oh','+ITJ'),('what','+WDET'),('and','+COORD'),('then','+ADV'),('we','+PRONPERS'),('have','+VHPRES'),('to','+INFTO'),('unload','+VINF'),('right','+ADJ')],
- [('I','+PRONPERS'),('think','+VPRES'),('it','+PRONPERS'),('is','+VBPRES'),('think','+VINF'),('that','+PRON'),('is','+VBPRES'),('gonna','+guessed+NOUN'),('add','+VPRES'),('up','+ADV'),('to','+PREP'),('fourteen','+CARD'),('let','+VPAST'),('us','+PRONPERS'),('see','+VPRES'),('um','+guessed+ADJ'),('avon','+g

uessed+ADJ'),('to','+INFTO'),('pick','+VINF'),('up','+ADV'),('the','+DET'),('three','+CARD'),('and','+COORD'),('get','+NOUN'),('back','+ADV'),('to','+PREP'),('avon','+guessed+NOUN'),('is','+VBPRES'),('six','+CARD'),('right','+ADJ'),('to','+INFTO'),('load','+VINF'),('is','+VBPRES'),('is','+VBPRES'),('seven','+CARD')],

- [('because','+COSUB'),('we','+PRONPERS'),('have','+VHPRES'),('to','+INFTO'),('mm','+MEAS'),('maybe','+ADV'),('maybe','+ADV'),('we','+PRONPERS'),('try','+VPRES'),('like','+PREP'),('taking','+PARTPRES'),('two','+CARD'),('boxcars','+NOUN'),('of','+PREP'),('well','+NOUN'),('taking','+PARTPRES'),('taking','+NOUNING'),('one','+CARDONE'),('boxcar','+NOUN'),('would','+VAUX'),('be','+VBINF'),('sufficient','+ADJ')],
- [('when','+WADV'),('and','+COORD'),('if','+COSUB'),('we','+PRONPERS'),('took','+VPAST'),('or','+COORD'),('no','+DET'),('wait','+NOUN'),('here','+ADV'),('here','+ADV'),('we','+PRONPERS'),('go','+VPRES')],
- [('if','+COSUB'),('if','+COSUB'),('we','+PRONPERS'),('take','+VPRES'),('um','+guessed+NOUN'),('let','+VPRES'),('us','+PRONPERS'),('see','+VPRES'),('because','+COSUB'),('we','+PRONPERS'),('have','+VHPRES'),('to','+PREP'),('we','+PRONPERS'),('have','+VHPRES'),('to','+INFTO'),('get','+VINF'),('the','+DET'),('bananas','+NOUN'),('too','+ADV')],
- [('so','+ADV'),('this','+PRON'),('we','+PRONPERS'),('have','+VHPRES'),('to','+INFTO'),('get','+VINF'),('to','+PREP'),('avon','+guessed+ADJ'),('somehow','+ADV')],
- [('and','+COORD'),('then','+ADV'),('take','+VINF'),('those','+PRON'),('uh','+ITJ'),('to','+PREP'),('and','+COORD'),('then','+ADV'),('go','+VPRES'),('to','+PREP'),('dansville','+guessed+ADJ')],
- [('drop','+NOUN'),('off','+PREP'),('that','+DET'),('boxcar','+NOUN'),('and','+COORD'),('take','+VPRES'),('well','+ADV'),('yes','+VINF'),('drop','+NOUN'),('off','+PREP'),('the','+DET'),('boxcar','+NOUN'),('of','+PREP')],
- [('ok','+ADV'),('so','+ADV'),('elmira','+guessed+ADJ'),('to','+PREP'),('corning','+PARTPRES'),('is','+VBPRES'),('two','+CARD'),('and','+COORD'),('loading','+NOUNING'),('is','+VBPRES'),('three','+CARD'),('and','+COORD'),('to','+PREP'),('dansville','+guessed+NOUN'),('is','+VBPRES'),('four','+CARD')],
- [('ok','+ADJ'),('so','+COSUB'),('that','+PRON'),('is','+VBPRES'),('and','+COORD'),('then','+ADV'),('we','+PRONPERS'),('are','+VBPRES'),('done','+VDPAP'),('with','+PREP'),('the','+DET'),('oranges','+NOUN')],
- [('not','+NOT'),('yet','+ADV'),('ok','+ADV'),('four','+CARD'),('four','+CARD'),('boxcars','+NOUN'),('of','+PREP'),('oranges','+NOUN'),('to','+PREP'),('bath','+NOUN'),('so','+ADV')],
- [('uh','+ITJ'),('do','+VDPRES'),('they','+PRONPERS'),('only','+ADV'),('need','+VPRES'),('one','+CARDONE'),('engine','+NOUN'),('for','+PREP'),('both','+QUANT'),('of','+PREP')],

- [('um','+guessed+ADJ'),('the','+DET')],
- [('ok','+ADJ'),('um','+guessed+NOUN'),('they','+PRONPERS'),('requested','+VPAST'),('a','+DET'),('tanker','+NOUN'),('so','+ADV'),('how','+WADV'),('do','+VDPRES'),('we','+PRONPERS')],
- [('bath','+NOUN'),('and','+COORD'),('two','+CARD')],
- [('and','+COORD')],
- [('ok','+ADJ'),('so','+COSUB'),('you','+PRONPERS'),('wanted','+VPAST'),('to','+INFTO'),('do','+VDINF'),('what','+WPRON'),('again','+ADV'),('you','+PRONPERS'),('wanted','+VPAST'),('to','+INFTO')],
- [('and','+COORD'),('then','+ADV'),('you','+PRONPERS'),('want','+VPRES'),('to','+INFTO'),('go','+VINF'),('to','+INFTO')]
- [('ok','+ADJ'),('and','+COORD'),('then','+ADV'),('we','+PRONPERS'),('need','+VPRES'),('to','+PREP'),('we','+PRONPERS'),('are','+VBPRES'),('leaving','+VPROG'),('leaving','+VPROG'),('leaving','+VPROG'),('the','+DET'),('boxcars','+NOUN'),('and','+COORD'),('we','+PRONPERS'),('are','+VBPRES')],
- [('um','+guessed+NOUN'),('here','+ADV'),('is','+VBPRES'),('the','+DET'),('problem','+NOUN'),('we','+PRONPERS'),('need','+VAUX'),('to','+INFTO'),('transport','+VINF'),('two','+CARD'),('tankers','+NOUN'),('of','+PREP'),('orange','+ADJ'),('juice','+NOUN'),('to','+PREP'),('avon','+guessed+ADJ'),('and','+COORD'),('three','+CARD'),('boxcars','+NOUN'),('of','+PREP'),('bananas','+NOUN'),('to','+PREP'),('elmira','+guessed+ADJ'),('uh','+ITJ'),('the','+DET'),('bana','+guessed+ADJ')],
- [('and','+COORD'),('we','+PRONPERS'),('need','+VAUX'),('to','+INFTO'),('go','+VINF'),('to','+PREP'),('um','+guessed+ADJ'),('avon','+guessed+NOUN'),('and','+COORD')],
- [('can','+VAUX'),('we','+PRONPERS'),('use','+VPRES'),('the','+DET'),('boxcars','+NOUN'),('there','+ADV'),('in','+PREP'),('dansville','+guessed+ADJ'),('since','+COSUB'),('there','+ADV'),('is','+VBPRES'),('three','+CARD'),('there','+ADV')],
- [('ok','+ADJ'),('um','+guessed+NOUN'),('that','+PRONREL'),('would','+VAUX'),('be','+VBINF')],
- [('load','+NOUN'),('the','+DET'),('oranges','+NOUN'),('and','+COORD')],
- [('and','+COORD'),('then','+ADV'),('on','+ADV'),('to','+PREP'),('dansville','+guessed+ADJ'),('which-ll','+guessed+NOUN'),('take','+VPRES')],
- [('ok','+ADV'),('you','+PRONPERS'),('can','+VAUX'),('not','+NOT'),('ship','+VINF'),('them','+PRONPERS'),('at','+PREP'),('the','+DET'),('same','+ADJPRON'),('time','+NOUN'),('because','+COSUB'),('there','+ADV'),('is','+VBPRES'),('only','+ADV'),('one','+CARDONE'),('track','+NOUN')],

- [('oh','+NOUN'),('ok','+ADV'),('well','+ADV'),('then','+ADV'),('I','+PRONPERS'),('want','+VPRES'),('to','+INFTO'),('send','+VINF'),('one','+CARDONE'),('and','+COORD'),('then','+ADV'),('the','+DET'),('other','+ADJPRON'),('right','+ADJ'),('after','+COSUB'),('it','+PRONPERS')],
- [('ok','+ADJ')],
- [('um','+guessed+ADJ')],
- [('and','+COORD'),('then','+ADV'),('uh','+ITJ'),('I','+PRONPERS'),('want','+VPRES'),('to','+INFTO'),('send','+VINF'),('I','+PRONPERS'),('want','+VPRES'),('to','+INFTO'),('send','+VINF'),('uh','+ITJ'),('one','+CARDONE'),('boxcar','+NOUN'),('full','+ADV'),('of','+PREP'),('oranges','+NOUN'),('to','+PREP'),('bath','+NOUN'),('and','+COORD'),('then','+ADV'),('another','+DET'),('one','+CARDONE'),('right','+ADJ'),('after','+COSUB'),('it','+PRONPERS')],
- [('ok','+ADV'),('so','+ADV'),('say','+VINF'),('I','+PRONPERS'),('send','+VPRES'),('engine','+NOUN'),('e','+PROP'),('two','+CARD'),('to','+PREP'),('corning','+PARTPRES'),('that','+PRONREL'),('takes','+VPRES'),('two','+CARD'),('hours','+NOUN')],
- [('yes','+NOUN')],
- [('and','+COORD'),('then','+ADV'),('I','+PRONPERS'),('load','+VPRES'),('up','+ADV'),('that','+COSUB'),('boxcar','+NOUN'),('with','+PREP'),('oranges','+NOUN'),('and','+COORD'),('send','+VPRES'),('it','+PRONPERS'),('to','+PREP'),('bath','+NOUN'),('it','+PRONPERS'),('takes','+VPRES'),('an','+DET'),('hour','+NOUN'),('to','+INFTO'),('load','+VINF'),('so','+COSUB'),('it','+PRONPERS'),('will','+VPRES'),('um','+guessed+ADJ'),('get','+NOUN'),('to','+PREP'),('bath','+NOUN'),('at','+PREP'),('five','+CARD'),('a-m','+guessed+ADJ')],
- [('it','+PRONPERS'),('will','+VAUX'),('get','+VINF'),('to','+PREP'),('bath','+NOUN'),('at','+PREP'),('five','+CARD')],
- [('yes','+NOUN')],
- [('ok','+ADJ')],
- [('and','+COORD'),('then','+ADV'),('the','+DET'),('second','+ORD'),('boxcar','+NOUN'),('I','+PRONPERS'),('will','+VAUX'),('send','+VINF'),('e','+PROP'),('three','+CARD'),('um','+guessed+ADJ')],
- [('with','+PREP'),('the','+DET'),('boxcarto','+guessed+ADJ'),('corning','+NOUNING'),('and','+COORD'),('then','+ADV'),('you','+PRONPERS'),('say','+VPRES'),('you','+PRONPERS'),('want','+VPRES'),('to','+INFTO'),('load','+VINF'),('oranges','+NOUN'),('onto','+PREP'),('that','+DET'),('boxcar','+NOUN')],
- [('engine','+NOUN'),('two','+CARD'),('to','+PREP'),('corning','+PARTPRES'),('two','+CARD'),('hours','+NOUN')],

- [('ok','+ADJ'),('mm','+CARD'),('let','+VPAST'),('us','+PRONPERS'),('see','+VPRES'),('and','+COORD'),('then','+ADV'),('bring','+VINF'),('it','+PRONPERS'),('back','+ADV'),('to','+PREP'),('elmira','+guessed+ADJ')],
- [('another','+DET'),('two','+CARD'),('hours','+NOUN')],
- [('ok','+ADV'),('so','+ADV'),('um','+guessed+ADJ')],
- [('I','+PRONPERS'),('will','+VAUX'),('take','+VINF'),('engine','+NOUN'),('two','+CARD'),('to','+PREP'),('corning','+PARTPRES'),('get','+NOUN'),('a','+DET'),('tanker','+NOUN'),('and','+COORD'),('bring','+VPRES'),('it','+PRONPERS'),('back','+ADV'),('to','+PREP'),('elmira','+guessed+ADJ'),('to','+INFTO'),('load','+VINF'),('orange','+ADJ'),('juice','+NOUN')],
- [('and','+COORD'),('then','+ADV'),('back','+ADV'),('to','+PREP'),('corning','+PARTPRES'),('to','+PREP'),('dansville','+guessed+ADJ'),('and','+COORD'),('then','+ADV'),('to','+PREP'),('avon','+guessed+ADJ')],
- [('how','+WADV'),('long','+ADJ'),('will','+NOUN'),('that','+PRONREL'),('take','+VPRES'),('from','+PREP'),('elmira','+guessed+ADJ'),('to','+PREP'),('corning','+PARTPRES'),('and','+COORD'),('then','+ADV'),('to','+PREP'),('dansville','+guessed+ADJ'),('to','+PREP'),('avon','+guessed+ADJ')],
- [('ok','+ADJ'),('uh','+ITJ'),('just','+ADV'),('that','+COSUB'),('that','+DET'),('last','+ADJ'),('trip','+NOUN'),('from','+PREP'),('elmira','+guessed+ADJ'),('all','+QUANT'),('the','+DET'),('way','+NOUN'),('to','+PREP'),('avon','+guessed+ADJ')],
- [('mm-hm','+guessed+ADJ')],
- [('that','+PRONREL'),('will','+VAUX'),('take','+VINF'),('uh','+ITJ'),('two','+CARD'),('three','+CARD'),('four','+CARD'),('five','+CARD'),('six','+CARD'),('hours','+NOUN'),('in','+PREP'),('all','+QUANT'),('from','+PREP'),('elmira','+guessed+ADJ'),('to','+PREP'),('avon','+guessed+ADJ')],
- [('six','+CARD')],
- [('ok','+ADJ')]
- [('so','+ADV'),('that','+PRON'),('was','+VBPAST'),('four','+CARD'),('hours','+NOUN'),('for','+PREP'),('the','+DET'),('engine','+NOUN'),('to','+INFTO'),('get','+VINF'),('to','+PREP'),('corning','+PARTPRES'),('and','+COORD'),('then','+ADV'),('back','+ADV'),('to','+PREP'),('elmira','+guessed+ADJ')],
- [('two','+CARD'),('hours','+NOUN')],
- [('two','+CARD'),('hours','+NOUN'),('and','+COORD'),('then','+ADV'),('to','+PREP'),('dansville','+guessed+ADJ')],
- [('uh','+ITJ'),('another','+DET'),('hour','+NOUN')],

- [('and','+COORD'),('to','+INFTO'),('bring','+VINF'),('one','+CARDONE'),('boxcar','+NOUN')],
- [('ok','+ADJ')],
- [('back','+ADV'),('to','+PREP'),('corning','+PARTPRES')],
- [('uh','+ITJ'),('another','+DET'),('hour','+NOUN'),('so','+COSUB'),('we','+PRONPERS'),('are','+VBPRES'),('we','+PRONPERS'),('are','+VBPRES'),('at','+PREP'),('four','+CARD'),('hours','+NOUN')],
- [('four','+CARD'),('and','+COORD'),('then','+ADV'),('to','+INFTO'),('load','+VINF'),('the','+DET'),('boxcar','+NOUN'),('of','+PREP'),('or','+COORD'),('with','+PREP'),('oranges','+NOUN')],
- [('um','+guessed+ADJ'),('another','+DET'),('hour','+NOUN')],
- [('and','+COORD'),('that','+PRONREL'),('takes','+VPRES'),('no','+DET'),('time','+NOUN'),('to','+INFTO'),('bring','+VINF'),('a','+DET'),('tanker','+NOUN'),('along','+PREP')],
- [('no','+ADV'),('I','+PRONPERS'),('mean','+VPRES'),('uh','+ITJ'),('yes','+VINF'),('it','+PRONPERS'),('takes','+VPRES'),('no','+DET'),('time','+NOUN')],
- [('ok','+ADJ'),('and','+COORD'),('then','+ADV'),('bring','+VPRES'),('all','+QUANT'),('that','+PRON'),('from','+PREP'),('corning','+PARTPRES'),('to','+PREP'),('elmira','+guessed+ADJ')],
- [('that','+PRONREL'),('is','+VBPRES'),('another','+DET'),('two','+CARD'),('hours','+NOUN')],
- [('so','+ADV'),('what','+WDET'),('time','+NOUN'),('are','+VBPRES'),('we','+PRONPERS'),('at','+PREP'),('now','+ADV'),('when','+COSUB'),('we','+PRONPERS'),('arrive','+VPRES'),('at','+PREP'),('elmira','+guessed+ADJ'),
- [('uh','+ITJ'),('it','+PRONPERS'),('is','+VBPRES'),('seven','+CARD'),('a-m','+guessed+ADJ')],
- [('ok','+ADJ'),('so','+COSUB'),('that','+PRON'),('is','+VBPRES'),('right','+ADJ'),('um','+guessed+ADJ')],
- [('well','+ADV'),('it','+PRONPERS'),('takes','+VPRES'),('us','+PRONPERS'),('an','+DET'),('hour','+NOUN'),('to','+INFTO'),('unload','+VINF'),('but','+ADV'),('is','+VBPRES'),('that','+COSUB'),('I','+PRONPERS'),('I','+PRONPERS'),('if','+COSUB'),('that','+PRON'),('is','+VBPRES'),('ok','+ADJ')],
- [('mm-hm','+guessed+ADJ')],
- [('I','+PRONPERS'),('do','+VDPRES'),('not','+NOT'),('think','+VINF'),('there','+ADV'),('is','+VBPRES'),('any','+ADV'),('it','+PRONPERS'),('just','+ADV'),('says','+VPRES'),('that','+COSUB'),('they','+PRONPERS'),('are','+VBPRES'),('due','+ADJ'),('to','+INFTO'),('be','+VBINF'),('processed','+VPAP'),('at','+PREP'),('the','+DET'),('factory','+NOUN'),('in','+PREP'),('elmira','+guessed+ADJ'),('at','+PREP'),('seven','+CARD'),('seven','+CARD'),('a-m','+guessed+ADJ'),('sharp','+NOUN')],
- [('is','+VBPRES'),('that','+DET'),('alright','+guessed+ADJ'),('uh','+ITJ'),('I','+PRONPERS'),('guess','+VPRES'),('so','+ADV'),('yes','+VINF'),('I','+PRONPERS'),('think','+VPRES'),('so','+ADV')]

- [('ok','+ADJ'),('so','+COSUB'),('we','+PRONPERS'),('will','+VAUX'),('make','+VINF'),('the','+DET'),('orange','+ADJ'),('juice','+NOUN'),('there','+ADV'),('in','+PREP'),('elmira','+guessed+ADJ')],
- [('right','+ADJ')]
- [('so','+ADV'),('um','+guessed+ADJ'),('we','+PRONPERS'),('are','+VBPRES'),('at','+PREP'),('seven','+CARD'),('now','+ADV'),('how','+WADV'),('long','+ADV'),('will','+VAUX'),('it','+PRONPERS'),('take','+VPRES'),('to','+INFTO'),('bring','+VINF'),('that','+DET'),('tanker','+NOUN'),('to','+PREP'),('corning','+PARTPRES'),('to','+PREP'),('bath','+NOUN'),('to','+PREP'),('avon','+guessed+ADJ')]
- [('ok','+ADV'),('you','+PRONPERS'),('can','+VAUX'),('only','+ADV'),('carry','+VINF'),('um','+guessed+ADJ'),('you','+PRONPERS'),('can','+VAUX'),('carry','+VINF'),('three','+CARD'),('empty','+ADJ'),('boxcars','+NOUN'),('that','+PRONREL'),('would','+VAUX'),('be','+VBINF'),('fine','+ADJ'),('but','+COSUB'),('once','+ADV'),('you','+PRONPERS'),('get','+VPRES'),('them','+PRONPERS'),('loaded','+VPAST')]
- [('can','+VAUX'),('carry','+VINF'),('as','+PREPADVAS'),('many','+QUANT'),('um','+guessed+ADJ'),('unloaded','+ADJPAP'),('boxcars','+NOUN'),('as','+PREPADVAS'),('you','+PRONPERS'),('want','+VPRES'),('but','+COSUB'),('once','+ADV'),('you','+PRONPERS'),('get','+VPRES'),('them','+PRONPERS'),('loaded','+VPAST'),('you','+PRONPERS'),('can','+VAUX'),('only','+ADV'),('carry','+VINF'),('three','+CARD'),('at','+PREP'),('a','+DET'),('time','+NOUN'),('um','+guessed+NOUN'),('you','+PRONPERS'),('can','+VAUX'),('carry','+VINF'),('three','+CARD'),('empty','+ADJ'),('boxcars','+NOUN')]
- [('that','+PRONREL'),('would','+VAUX'),('be','+VBINF'),('fine','+ADJ'),('but','+COSUB'),('once','+ADV'),('you','+PRONPERS'),('get','+VPRES'),('them','+PRONPERS'),('loaded','+VPAST'),('you','+PRONPERS'),('can','+VAUX'),('carry','+VINF'),('as','+PREPADVAS'),('many','+QUANT'),('um','+guessed+ADJ'),('unloaded','+ADJPAP'),('boxcars','+NOUN'),('as','+PREPADVAS'),('you','+PRONPERS'),('want','+VPRES'),('but','+COSUB'),('once','+ADV'),('you','+PRONPERS'),('get','+VPRES'),('them','+PRONPERS'),('loaded','+VPAST'),('you','+PRONPERS'),('can','+VAUX'),('only','+ADV'),('carry','+VINF'),('three','+CARD'),('at','+PREP'),('a','+DET'),('time','+NOUN')]

2. Annexe 2: Exemple d'annotation des extragrammaticalités dans un dialogue du *Trains Corpus*

2.1 Annotation des faux départs et autocorrections

Voici à titre d'exemple, l'annotation des faux-départs et autocorrections du dialogue d93-23.2.

No cas	Enoncé	Sortie du tagger de Xerox	Phénomène	Patron
		oh oh +ITJ I I +PRONPERS can can +VPRES + + +PUNCT can can +VAUX I I +PRONPERS + + +PUNCT put put +VPRES more much +QUANTCMP more more +QUANTCMP than than +COTHAN		
	utt15 : : u : oh **I can + <sil> can I** + put more than one boxcar on an	one one +CARDONE boxcar boxcar +NOUN on on +PREP an an +DET		
61	engine	enginee enginee +guessed+ADJ	Auto-cor	M1M2 ED M2M1 (ED = sil)

62	utt28 : u : okay <sil> and then <sil> how long does it take to get **to <sil> from** Corning to Dansville	okay okay +VPRES and and +COORD then then +ADV how how +WADV long long +ADV does do +VDPRES it it +PRONPERS take take +VPRES to to +INFTO get get +VINF to to +PREP from from +PREP Corning corn +PARTPRES to to +PREP Dansville Dansville +PROP	Auto-cor	R1 ED R1 (ED = sil)
63	utt58 : okay <sil> **and then I want to send those** two tankers back to Corning	okay okay +VPRES and and +COORD then then +ADV I I +PRONPERS want want +VPRES to to +INFTO send send +VINF **that that +PRON** and and +COORD then then +ADV I I +PRONPERS want want +VPRES to to +INFTO send send +VINF **those those +DET** two two +CARD tankers tanker +NOUN back back +ADV to to +PREP Corning corn +PARTPRES	Auto-cor	M1M2M3M4M5R1 ED M1M2M3M4M5R1' (R1 = pron; R1' =Det, ED = sil)

		um um +guessed+ADJ it it +PRONPERS takes take +VPRES six six +CARD hours hour +NOUN to to +INFTO get get +VINF from from +PREP Dansville Dansville +PROP to to +PREP Avon Avon +PROP so so +ADV that that +ADV 's 's +POSS i- i- +open+ADJ		
64	utt86: um <sil> **it takes six hours <sil> to get from Dansville <sil> to Avon <sil> so that's** <sil> ɨ it takes three hours so that's um six a.m.	it it +PRONPERS takes take +VPRES three three +CARD hours hour +NOUN so so +ADV that that +ADV 's 's +POSS um um +guessed+ADJ six six +CARD a.m. a.m. +ADV	Faux-dép	

2.2 Annotation des répétitions

Voici l'annotation des répétitions dans le dialogue d93-23.2 :

Enoncés avec répétitions simples	**Patron**	**Enoncés avec répétitions avec édition**	**Patron**
utt22: u: **no no no** leave the boxcars at Corning **(imbriquée)**	M1M1M1	utt10: u: oh it's midnight <sil> okay <brth> uh <sil> okay <sil> then	M1 ED M1 (ED = bruit)
utt24: u: take one engine **and <sil> and and** two <sil> and two tankers and <sil> put it take it to Elmira **(imbriquée)**	M1M1EDM1 (ED= sil)	utt17: uh **you can put <u><sil> wait</u> <sil> you can put** <sil> three <sil> loaded boxcars <sil> or tanker cars <sil> on an engine <sil> and <sil> any number of unloaded cars	M3 ED M3 (ED = sil + wait + sil)
utt44: right <sil> and then <sil> right **go go** from Corning to Dansville	M1M1	utt18: u: oh <sil> **oh okay <sil> oh okay <sil> well <sil> okay then I want to uh <sil> then I want to** <sil> take one of the engines	M1M2 ED1 M2M1 ED2 M1 M3 ED3 M3 (ED1 = sil) (ED2 = sil) (ED3 = uh sil then)
utt77: so <sil> you sent <sil> **engine <sil> engine** E two <sil> and engine <sil> E three <sil> to <sil> Corning		utt32: u: okay <brth> then <sil> I want to send <sil> um <sil> one engine[61]	M1M1 M1 M1
		utt33: and <sil> one <sil> and and and <sil>	M1M1, M1M1
		utt80: you s- at the same time you sent <sil> um <sil> E two <sil> uh with <sil> with three tankers <sil> to <sil> Elmira	M1 M1

[61] Tour ajouté pour garder le contexte dialogique. Dans ce cas, ce la permet de vérifier qu'il ne s'agit pas d'un comptage (one, two, three, four, etc.).

Dans ce dialogue, nous avons trouvé un cas de surgénération potentille. Il est présenté dans le tableau suivant :

No cas	Énoncé	Sortie du tagger de Xerox	Remarques
37	utt56: s: + yes + <sil> + **they + are**	yes yes +NOUN + + +PUNCT + + +PUNCT they they +PRONPERS + + +PUNCT are be +VBPRES	*They are*, dans cet exemple, est une ellipse bien formée. Il ne s'agit pas d'une incomplétude.

3. Annexe 3 : exemples de règles syntaxiques utilisées pour le traitement des faux-départs

- **Méta règle pour un faux départ :**

faux_dep_sn_verb_ed → frontière_début chunk_inc_sn_verb édition phrase_déc

- **Règle de la frontière de début d'une extragrammaticalité :**

frontière_début → marque_de_debut

..

..

- **Quelques règles des marques de début :**

marque_de_debut → cosub[62] /* Conjonction de subordination */

marque_de_debut → hésitation

..

..

- **Règle d'incomplétude d'un chunk :**

Chunk_inc_sn_verb → verb_pres_past_infinitf not*[63]:

verb_pres_past_infinitf → verb_inf_trans /* Verbe transitif à l'infinitif */

verb_pres_past_infinitf → verb_pres_trans /* Verbe transitif au présent */

..

..

[62] Comme l'entrée de l'analyseur est une série de mots pré-tagués nous n'avons pas de règles dont la partie gauche est un terminal.

[63] L'étoile signifie qu'un élément est facultatif.

- **Règle de la zone d'édition :**

édition → mot_édition

mot_édition → hésitation

mot_édition → adv. /* adverbe*/

..

..

- **Règle du contexte droit :**

Phrase_déclarative → synt_nominal synt_verbal

Synt_nominal → sn_pron

..

..

Synt_verbal → sv_adj

..

..

Sn_pron → pronpers adv* /* syntagme pronominal adverbial */

Sv_adj → verb_pres adj verb_inf

4. Annexe 4 : Annotation du corpus de réservation hôtelière

Voici, à titre d'exemple, l'annotation de deux catégories conceptuelles correspondants à des arbres élémentaires dans le formalisme S-TSG.

1. Formule de demande :

- Auriez vous →[obj. Ch./disponible] →[date]→[svp].
- Je voudrais →[obj. Réserver] →[nb. Ch], [nb. personnes].
- J'aimerais/j'aurais voulu → [obj. Savoir/dispo. ch.] →[durée].
- Je voudrais →[obj. Une chambre].
- Je vous appelle pour → [obj. réserver]→[une chambre] →[date].
- J'aurais voulu →[obj.une chambre]→[date] →[nb. personnes].
- J'aimerais →[réserver]→[une chambre]→[carac-chambre]
- Je voudrais →[une chambre].
- Je voudrais →[réserver]→[date].
- J'aurais voulu →[..] // (C←dem-dispo-ch. /→H[rép. non disp +date]).
- Je désirerais →[une chambre].
- J'aurais voulu →[savoir]→[si vous auriez].
- Je vous appelle pour →[rés] →[une chambre].
- C'est pour [une réservation].
- Nous aurions besoin de →[nb. Ch.] →[carac-Ch.].
- Je souhaiterai → [réserver]→[une nuit]→[c'est possible] .
- Je souhaiterai →[avoir].
- Serait-il possible de →[réserver] ou [savoir (si)].
- Il me faudrait →[nb. Ch.]→ [carac-Ch.].
- C'est possible ?←[f-dem]←[réserver]← [nb. Ch].
- Si c'est possible ← [obj-demdé].
- Ce serait pour → [une réservation] → [date].

- Que pouvez vous m'offrir dans le genre →[carac-Ch.].
- C'est juste pour → [un renseignement].
- Je voulais → [savoir].
- Nous prendrons → [carac-ch]→[svp].
- J'aimerais →[louer une chambre]→[date].
- J'ai téléphoné il y a une heure pour →[réserver une chambre pour ce week-end].
- Je vous téléphone pour → [réserver une chambre].
- J'aimerais →[des renseignements].
- Auriez vous → [une Ch] → [date + durée].
- Puis je avoir→[un Ch.]→ [avec parking].
- Pourrais-je → [la voir].

2. Heure arrivée :

- Soir.
- Au début d'après midi.
- 22 heures/ C← j'arriverais.
- au soir / C← jour.
- Très tard (..) vers 22 heures ←/ [j'arriverai].
- Le matin C← [j'arriverai].
- Enfin de matinée. H← dem heure arrivée.
- Vers 22 heures.
- Tard C←exp. Arrivée. → au train de 22 Heures.
- Un peu tard, vers 23 H. ←[f-arriv].
- A 23 pas plus tard←/H[heure fermeture du restaurant].
- A minuit trente←[mon train arrive en gare].
- à midi. [Probablement].
- Mais avant 19 heures de toute façons.
- En fin d'après midi .←[nous arivons].

5. Annexe 5 : le corpus initial ainsi qu'un exemple d'énoncés dérivés utilisés lors de la campagne d'évaluation par défi

5.1 Le corpus initial

<1> bon dans ces conditions alors réservez moi une chambre sympa et calme surtout pour le 6 février prochain </1>

<2> mon train arrive le 10 Décembre à 19 heures 37 </2>

<3> pardon encore une chose quel est le prix de la chambre </3>

<4> euh le 8 octobre je voudrais une baignoire si c'est possible </4>

<5> je voudrais une simple plutôt calme si c'est possible </5>

<6> bonjour madame je viens de la part de l'office du tourisme il paraît que vous avez encore des chambres libres </6>

<7> à partir de ce soir pour trois nuits quels types de chambres avez vous </7>

<8> bien je vous remercie à lundi donc j'arriverai vers 18 heures 30 </8>

<9> c'est parfait pouvez vous la réserver du 13 au 16 août au nom de gaud </9>

<10> c'est parfait vous me la réservez mais j'arriverai tard au train de 22 heures qui arrive de yon je crois </10>

<11> et ehh j'étais assez intéressé par voir tout ce qui est euh nature les animaux la faune et la flore quoi</11>

<12> le parc de stelvio c'est un parc national qui se trouve à l'ouest du trentino il y de très belles montagnes là bas </12>

<13> d'accord par contre je voulais savoir aussi s'il y a la possibilité de euh de louer les accessoires pour faire du ski </13>

<14> est-ce qu'il y a une salle de jeux si des fois il faisait pas très beau</14>

<15> à cavalese il y a des initiatives pour les enfants et à l'hôtel on va voir s'il vous plaît non je regrette à l'hôtel il n'y a pas d'animation</15>

<16> j'appelle pour avoir des renseignements pour organiser une semaine de ski au mois de décembre</16>

<17> je pense partir avec mes enfants et ma femme et donc j'aimerais avoir des renseignments sur la meilleure ville quel est le meilleure endroit où aller </17>

<18> vous désirez une région où on peut skier avec des enfants qui sont débutants</18>

<19> il y a une bonne station à cavalese avec la possibilité de faire des excursions de l'alpinisme et il y a aussi la possibilité de voir plusieurs animaux</19>

<20> j'aimerais un endroit où je puisse faire du ski de fond et du ski de piste les deux quoi</20>

5.2 Un extrait du corpus dérivé

Voici les quinze énoncés générés par notre collègue Jeanne Villaneau correspondant au premier énoncé initial :

<1.1> pour le 26 février prochain vous pouvez me réserver une chambre sympa et calme s il vous plaît </1.1>

<1.2> est ce que vous auriez pu me réserver une chambre calme pour le 26 février prochain </1.2>

<1.3> bon dans ces conditions alors réservez moi une chambre sympa et calme surtout pour les 26 et 27 février prochains </1.3>

<1.4> j aimerais que vous me réserviez une chambre pour le 26 février prochain et calme surtout</1.4>

<1.5> pour le 26 février prochain réservez moi une chambre bien calme surtout </1.5>

<1.6> j aimerais une chambre sympa et calme surtout pour le 26 février prochain est ce que vous pouvez m en réserver une </1.6>

<1.7> bon réservez moi pour le 26 février prochain une chambre calme si possible </1.7>

<1.8> est ce qu il y a possibilité de réserver une chambre sympa et calme surtout pour le 26 février prochain </1.8>

<1.9> je cherche une chambre sympa et calme surtout pouvez vous m en réserver une pour le 26 février prochain </1.9>

<1.10> réservez moi une chambre pour le 26 février prochain sympa et calme surtout </1.10>

<1.11> bon dans ces conditions alors réservez moi une chambre sympa et calme surtout pour le 26 non pardon pour le 27 février </1.11>

<1.12> pour le 26 février pouvez vous me réserver une chambre qui serait à la fois bien calme et agréable </1.12>

<1.13> réservez moi une chambre pour le 26 février prochain sympa et calme la chambre surtout </1.13>

<1.14> je devrais arriver le 26 février prochain réservez moi une chambre sympa et calme surtout </1.14>

<1.15> une chambre sympa et calme surtout vous me la réservez s il vous plaît pour le 26 février </1.15>

6. Annexe 6 : Description de la méthode DCR étendue

Voici la description de la méthode DCR étendue effectuée dans notre article qui a été publié dans les actes de la conference LREC02, 29-31 Mai, Las Palmas.

Toward an objective and generic Method for Spoken Language Understanding Systems Evaluation: an extension of the DCR method

Mohamed-Zakaria KURDI[1 and 2]

[1]Natural Interactive Systems Laboratory (NISLab)
University of Southern Denmark
Main campus: Odense University
Science Park 10
DK-5230 Odense M, Denmark
Kurdi@nis.sdu.dk
http://www.nis.sdu.dk/~kurdi/

Mohamed AHAFHAF[2]

[2]Laboratoire CLIPS – IMAG
(GEOD)
BP. 53
380401, Grenoble cedex 09, France
mohamed.ahafhaf@imag.fr

Abstract

In this paper, we present an extension of the DCR method, which is a framework for the deep evaluation of Spoken Language Understanding (SLU) Systems. The key point of our contribution is the use of a linguistic typology in order to generate an evaluation corpus that covers a significant number of the linguistic phenomena we want to evaluate our system on. This allows to have more objective and deep evaluation of SLU systems.

1. Introduction

During the last decade, there was an increased interest in spoken language dialogue systems and especially in their Spoken Language Understanding (SLU) components. Many approaches of spoken language with different theoretical backgrounds were proposed and implemented. This necessitated the development of different evaluation methodologies in order to test the effectiveness of these different approaches. The main common methodologies are quantitative ones like the ATIS evaluation campaign in which the performance of the tested system is measured by comparing its real output with a corresponding analysis by hand.
Despite their interest, these methods do not provide a detailed diagnostic of the negative and positive aspects of the system in term of linguistic phenomena processing. Further more, they require a lot of adaptations (precise task, system's output format, etc.) in order to make an objective comparison between different systems.
To avoid the limitations of quantitative methods, several deep schemes were proposed. Among these schemes, the DCR (Declaration, Control, Reference) method seems the most ambitious to provide a general framework for a qualitative evaluation of spoken language systems (Zeiliger et al., 1997), (Antoine et al., 1998). Despite the improvement of the evaluation quality with this method, it lacks of systematicity that makes the comparison of the results of different systems hard to do. In this paper we present an extension of the DCR method that allow to provide both deep and systematic evaluation.
The outline of this paper is as follows: in section two we present the major requirements of an objective evaluatin method of a SLU system. In section three, we present the main aspects of the DCR method. Our method is described in section four. In section five we provide a description of our

experiments and results and finally conclusion and perspectives will close the paper.

2. Major requirements from an objective evaluation method of SLU systems

The major requirements of an objective and generic method for evaluating SLU systems are:

- **Task independence:** the method should be applied to different systems whatever are their tasks.
- **Output format independence and analysis level independence:** one of the major problems that face a generic evaluation method is to be able to compare systems with different output formats or to test systems with different analysis level (syntactic parsing or semantic analysis).
- **Predictivity:** the method should provide a detailed diagnosis of the errors of the system. This allows to drive future improvements of the system.
- **Objectivity:** the evaluation corpus should contain representative linguistic phenomena of the language it is designed to process.
- **Flexibility:** partial evaluation should be possible. For example, one should be able to evaluate his system on a specific phenomenon or a small set of phenomena that he consider as particularly interesting for his system.

3. Presentation of the DCR method

The DCR method was proposed as an attempt to satisfy the major part of the requirement presented above. It is based on the generation of derived test sentences on the basis of initial ones extracted from the corpus on which the system is built. The derived corpus contains a set of groups where every group is dedicated to the evaluation of a unique linguistic phenomenon. Every DCR test consists of three components (Antoine et al., 2000):

1. The Declaration **D**: it corresponds to an ordinary utterance that may be uttered by the system's users.
2. The Control **C**: it consists of a modified version of the utterance D usually with a focus on a precise phenomenon that is present in D.
3. The Reference **R**: it consists of a Boolean value which accounts for the coherence of the utterances C and D.

Here is an example of the DCR test:
<D> I want a double room with with Internet uh Internet connection
<C> I want a double room
<R> False
The main problem of this method is that it does not provide a linguistic framework for the derivation of the D utterances (initial utterances) into C utterances (derived utterances). In fact, the derived utterances are generated following quasi-subjective and task dependent criteria without any guaranty of production systematicity. This makes the comparison of the results of two different systems with different application domains very hard to do.

4. Presentation of our method

In order to overcome the systematicity and derivation objectivity problems in the DCR method, we propose an extended version of it that allows to generate the derived utterances following an a priori defined linguistic typology. The key features of our method are presented in the following paragraphs:

4.1. Initial corpus

The initial corpus consists of a set of utterances relevant to the task of the system. These utterances are chosen following two criteria: in one hand, they have to cover the different semantic aspects of the system and in the other hand, they should provide a riche syntactic base for the derivation operations (they should contain different syntactic structures).

4.2. The derivation grammar

The derivation grammar is built on the basis of syntactic typology that has two main resources:

1. **Existing grammars:** the existing classical grammars and linguistic typological descriptions of the language of the system we want to evaluate are valuable source for the creation of the derivation grammar. They are particularly important because they provide a general and almost exhaustive description of the different standard syntactic phenomena.
2. **Existing linguistic resources:** spoken language corpora are analysed in order to extract the occurrences of different forms of the phenomena we want to test. The major motivation of extracting a part of our rules directly from these corpora is to take into consideration the linguistic phenomena of spoken language that are not systematically considered in the classical grammar books and linguistic typological studies (since they are mainly concerned with written language rather than spoken one).

The transformation grammar contains a set of rules divided into subgroups containing each the set of rules specialized in a specific linguistic phenomena. The rules are written with the following format:

1. **Rules** - two rules are given: the rule corresponding to the structure of the element in the initial utterance on which we want to apply the derivation. This rule is given only

when the derivation is applied on a complex structure. The second rule concerns the transformation to be applied.

2. **Transformation type** - we distinguished between two types of transformations:
   - **a.** Internal transformations: they consist of a systematic replacement of some elements inside the test units.
   - **b.** External transformations: they consist of making some operations at the global level of the utterance: by deleting some units, changing their position, etc.
3. **Application conditions** - each derivation rule is associated to a set of application conditions. These conditions are intended to make it precise the nature of test unit to which this transformation operation may be applied. This may lead the human generator in one hand to be systematic in applying the transformations to the whole units to which it might be applied and in the other hand that allows to avoid the generation of agrammatical or semantically inconsistent utterances (especially if the generation is done by a non native speaker).

Two examples of derivations rules with their application conditions are presented below:

1. **An example of an internal transformation rule:**

Rule: Sn (sp)[64] ® pas Sn
[NP (PP) ® not NP]
Type: Intra-unit derivation.
Application conditions: this rule may be applied to each non-pronominal Sn (NP) in an elliptic context. For example it cannot be applied to the Sn *une chambre* (a room) in a context such: *je voudrais réserver une chambre* (I want to reserve a room)[65].
Example:
This rule may be applied to the elliptical utterance: *une chambre* (a room) which becomes after the transformation: *pas une chambre* (not a room).

2. **An example of an external transformation rule:**

Rule: Sn Sp ® Sp Sn
[NP PP ® PP NP]
Type: inter-unit derivation.
Application conditions: this rule may be applied to any type of Sn and Sp.
Example: the utterance: *une chambre pour deux personnes* (a room for two persons) becomes after the derivation: *pour deux personnes une chambre*.

4.3. Derived corpus

The derived corpus is obtained after applying methodologically the transformations operations defined in the derivation grammar to the initial corpus. Contrary to the DCR procedure, the derivation is done by applying a set of predefined transformations on the basic units in the utterance.

4.3.1. Test unit

One of the main weaknesses in the DCR method is that it does not use an objectively predefined method for the segmentation of the input utterance in order to extract the basic units of evaluation. The

[64] The elements between brackets are alternatives to the previous ones.

[65] In order to give an idea about the syntactic changes we are giving literal translation of the examples.

segmentation of the initial utterance is done following communicative criteria as we proposed for our formalism Sm-TAG (Kurdi, 2001).

Each evaluation unit corresponds to a unique conceptual segment. A conceptual segment is a set (chunk) of words playing a particular semantic/pragmatic role in the utterance. These roles involve a great variety of cognitive and linguistic considerations such that (Androws, 1985):

- **Topicality of the utterance:** in topic comment articulation, some chunks play usually the role of the topic, which indicates what the utterance is about. The comment, which is the remainder of the sentence, provides information about the topic.
- **Given vs. Non-given:** what the system is presumed to know *a priori* (via the task model) vs. what it doesn't know.
- **Importance:** what is forwarded as important vs. what is backwarded as secondary.
- **Specificity:** whether the speaker is referring to a particular instance of an entity or to this entity in itself.

For example, the utterance: *Je voudrais réserver une chambre pour deux personnes* is segmented in the following way with our segmentation criteria:
[je voudrais (topic1)] [réserver (comment1)] [une chambre (comment2/topic2)] [pour deux personnes (comment3)]

The main motivation of using these discourse based rather than classical syntactic phrase based units is that this allows us to reduce the number of derivation and to focus mainly on the syntactic transformations that has a significant implication on semantic and pragmatic interpretation of the utterance.

4.3.2. The derivation process

The derivation process consist of transforming the initial utterances into derived ones by mean of the generation rules. As we saw, the generation rules contain a set of general guidelines for the grammar generator in order to avoid overgeneration and other generation problems. The first step in the generation is the segmentation of the initial utterances following the criteria presented in the 4.3.1. Paragraph. The second step consists of applying systematically the whole transformations described in the derivation grammar to the evaluation units that we obtained after the segmentation of the initial utterances. In order to change only one variable at time, each derived utterance consists of the transformed unit plus the rest of the utterance (without any change) except if the derivation described by a specific rule requires the deletion of a part of the utterance. For example, let us take the following initial utterance: *Je voudrais réserver une chambre pour deux personnes*, and the following derivation rule:
verbe ® ne verbe pas [verb ® pre-negation mark verb post-negation mark]

The previous rule might be applied only to the first unit (since it is the only unit in the utterance with a verbal head). Although the result of the application of the rule is a well formed utterance: *je ne voudrais pas*, the generated utterance is ***je ne voudrais pas*** *réserver une chambre pour deux personnes* since the derivation rule does not require the deletion of any element in the utterance.
In the other hand, if we have a derivation rule such:
Sn Sv Sn ® Sn [NP VP NP ® NP]
The derived utterance will contain only one Sn (NP) since the deletion of the rest of the elements is a part of the derivation itself.

5. The experiments

5.1. The Oasis system

As a first experiment of our methodology, we choose to make a test of the Oasis system (Kurdi, 2001). This system is based on the Semantic Tree Association Grammar Sm-TAG which is a hybrid formalism combining both syntactic and semantic information in one framework. The general architecture of this system is presented in the following figure:

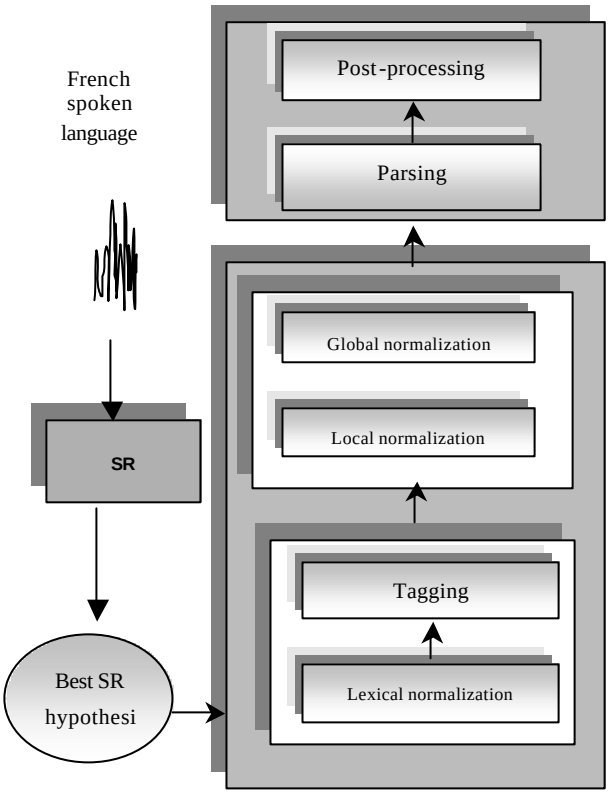


Figure 1. The architecture of our system

As we can see in the previous figure, Oasis system is based on a serial architecture containing 6 modules divided into three main parts from functional point of view:

1. **Pre-processing:** the pre-processing is mainly based on pattern matching techniques and it is intended to correct lexical extragrammaticalities, self-corrections and repetitions.
2. **Parsing:** we are using a 4 step parsing algorithm based on the combination of inductive rules to Recursive Transition Networks RTNs. The key property of this algorithm is the use of partial and selective parsing approach that allows the system to detect and process the relevant parts of the utterance.
3. Post-processing: we have a post-processing module based on semantic meta-rules intended to normalise the false-starts.

5.2. The considered phenomena

We made an evaluation of this system on three syntactic phenomena that we considered as the particularly relevant for SLU systems. These phenomena are: negation, ellipsis, and extraction.

5.2.1. Negation

Negation is a multidimensional phenomenon that has at the same time lexical, grammatical, and semantic aspects. So, The negation phenomenon is not only a lexical or syntactic reality but also a semantic one. This is one of the main reasons for which we choose the negation as a phenomenon to test our system on. Moreover, in French, there are some interesting differences of negation use between spoken language and written language. For example, the word *ne* (one of the two negation adverbs in French) is often neglected in the informal spoken language like in *je réserve pas* (I reserve not) instead of *je ne reserve pas* in written language and formal spoken language.

We distinguished between three types of negation:

- **Verbal:** when the negation is about a verbal phrase like *je ne voudrais pas une chambre simple* (I do not want a single room).
- **Nominal and prepositional:** it concerns the negations of a nominal or prepositional phrase like: ***pas*** *une chambre* (not a room), ***pas*** *pour une personne* (not for one person) (this case is hybrid one: it combines the negation to the ellipsis).
- **Pronominal:** we can have cases like the utterance *rien* (nothing), *aucun* (nobody) (this case is a hybrid one: it combines the negation to the ellipsis).

5.2.2. Ellipsis

The ellipse phenomenon consists of the deletion of one element or more from the utterance without affecting its grammaticality and interpretability. Two major types of ellipsis may be distinguished: grammatical or contextual ellipsis.

- The grammatical ellipsis consists of deleting some words following pure syntactic criteria. For example, in a sentence such *réserves pas* (reserve not) the word *tu* (you) that has the subject function is deleted from the utterance.
- The contextual ellipsis are used frequently in dialogue context in order to avoid the repetition of the already said elements of the utterance. If we consider *je réserve pour demain* at the time of reservation with an agent, this one will understand the request

referring to both discourse context and domain of request (ticket, room, etc).

From syntactic point of view, we distinguished between two forms of ellipsis:

- Phrase ellipsis: consist of the deletion of one or more (nominal, verbal or prepositional) phrase from the utterance. For example, *une chambre* (a room) is an elliptical utterance from which the verbal phrase *je voudrais* is deleted.
- Word ellipsis: word ellipsis consists of the deletion of a word playing a specific role in a particular phrase. This word may be the head of the phrase (like the noun in a nominal phrase) or a normal element in it (like a determinant in a nominal phrase). For example, we may have an utterance such *deux* (two), where the noun (which is the head of the phrase) is deleted. In the other hand, we may have an utterance like *chambre simple* (room simple) where the determinant is deleted.

5.2.3. Extraction

The extraction is a phenomenon that allows displacing a phrase (usually prepositional phrase and adverbs) to the right or left of the adjacent phrase without affecting the meaning of the utterance. For example, the adverbial phrase *le 10 décembre à 19 heures 37* (the December 10th at 19 o'clock) in the utterance: *mon train arrive le 10 décembre à 19 heures 37* (my train arrive the December 10th at 19 o'clock) may be displaced to the beginning of the utterance and the transformed utterance becomes: *le 10 décembre à 19 heures 37mon train arrive*. The extraction's effect is to divide a sentence into two parts, sometimes on three parts depending on its size and constituents. The extraction is considered as a part of a wide problematic of the words order (Blasco-Dulbecco, 1999) in which we notice the apparition of others phenomenon as double-marking (double-marquage) (Benveniste, 1990) used frequently in spoken language.

We distinguished between different forms of the extraction following the position of the extracted element (preposition or postposition) as well as following the nature of the extracted elements (prepositional phrase, adverb, etc.)

5.3. The generation grammar and derived corpus

We used different grammatical sources in order to write the grammar. These sources include many grammar books like (Gadet, 1989), (Gadet, 1992), and linguistic typological studies like (Benveniste, 1997), (Blasco-Dulbecco, 1999). We also used three spoken language corpora: hotel reservation corpus (Hollard, 1997), Dali project corpus (Sabah, 1997), and Murol corpus (Caelen et al, 1997).

We obtained a total of 154 rules with: 105 negation rules, 17 ellipsis rules, and 32 extraction rules. Some of the rules are hybrid ones (they apply for two phenomena at the same time). These rules cover about 23% of the total number of derivation rules. In order to avoid double generation and allow the independence of the grammar of each phenomenon, the hybrid rules are labelled in a special way in the grammar sets.

In order to limit the number of generated utterances for this first experiment, we generated from one to three utterances corresponding to each rule. The multiple generations were done when we considered that the lexical change might have an effect on the behaviour of the system. Thus, we obtained 252 derived utterances on the basis of ten initial ones.

5.4. Evaluation results

Before we present the results of our evaluations, we resolved two issues:

Selective strategy effect: as we said in a previous section, our parsing algorithm is based on a selective strategy that allows it to detect the relevant part in the utterance. This leaded us to distinguish between two types of generated utterances: relevant utterances and irrelevant utterances. The difference between these two types is that in the relevant utterances the transformation described in the derivation rule is realized in an area relevant for the system (the utterance is then considered as relevant) or irrelevant for the system (the utterance is then considered as irrelevant). Only the relevant utterances were considered in the results calculation.

In the other hand, we considered only the assessed phenomena are considered in our evaluation except if there is an error with the processing of an irrelevant phenomena that was directly caused by a derivation. This limitation allows us to get concentrated only on our targeted phenomena rather than covering the rest.

Following our statistics, 27,8% of the generated utterances was irrelevant to the task of our system. In the other hand, 88,6% of the relevant cases was processed correctly. In only 2,5% of the cases the derivations caused an external error (an analysis error in a non targeted phenomenon). Following our analysis we found that 77,78% of the parsing errors are due to the undergeneration of the grammar while the 22,22% are due to the way in which some rules are implemented.

Below are presented the detailed results sorted by phenomenon.

5.4.1. Negations results

We obtained 157 utterances with negation. The results of the Oasis system on these utterances are presented in the following table:

Type of negation	% of the correctly processed cases
Verbal	91,66
Nominal and prepositional	84,61
Pronominal	78,57
Hybrid with extraction	-
Hybrid with ellipsis	81,59
Total	84,48

Table 1. Our results on the negation cases

As we can see in the previous table, Oasis system was able to process more easily the classical negation form (the verbal) than the less classical ones, especially the adverbial ones that requires in some cases a higher level of knowledge.

5.4.2. Ellipsis results

Our corpus contains 50 utterances with ellipsis cases. Our evaluation results on these utterances are presented in the following table:

Type of ellipsis	% of the correctly processed cases
Verbal phrase ellipsis	75
Nominal phrase ellipsis	100
Noun ellipsis	0
Determinant ellipsis	71,4
Hybrid: different forms of ellipsis with extraction	100
Total	**76,19**

Table 2. Presentation of our evaluation results on the ellipsis cases

As we can see in the above table, the Oasis system processing capacity varies following the degree of difficulty of the ellipsis cases. Its capacities are perfect in processing the classical nominal ellipsis cases. Concerning the verbal ellipsis it achieves a coverage of about 75% of the cases. In the case of noun ellipsis, we can see that the Oasis system has a null capacity of processing. This is due to the fact that this kind of ellipsis requires the knowledge of the dialogue context (which beyond the knowledge sources of Oasis) in which this elliptic utterance is realized.

5.4.3. Extractions results

We have 50 utterances with extractions. The results of our evaluation on these cases are presented in the following table:

Type of extraction	% of the correctly processed cases
Preposition	95,45
Postposition	94,54
Verbal	92,72
Nominal and prepositional	96,36
Adverbial	94,44
Total	94,11

Table 3. Our results on the extraction cases

Our results show that the position of extraction (preposition and postposition) has no real significance for the processing. In the other hand, it shows that the extractions of different constituents are processed in almost the same way although some of them are less frequently observed in spoken language corpora than the rest (like the verbal extractions).

6. Conclusion

In this paper, we presented an extension of the DCR methodology. The main motivations of our extension are:

1. To allow a systematic (and by consequent more objective) generation of the evaluation corpus.
2. To have a more deep diagnostic of the evaluated system.

For satisfying these two conditions, we defined a derivation method that allows to obtain an evaluation corpus build following an a priori defined linguistic typology of the phenomena we want to assess our system on. As we saw, this methodology is task and lexicon independent and allow to evaluate any system independently of the representation level of its output (syntactic, semantic or pragmatic representation).

The application of our method on the evaluation of an SLU system showed that it is realistic and that it allows to obtain a deep diagnostic of the reasons of success and failure of the system.

As a perspective of our work, we intend to apply our method to more than one SLU system (preferably with different approaches) in order to show that it may be used to compare not only the

involved systems but also the effectiveness of their approaches to the SLU task.
Finally, we are investigating the possibility of extending our methodology to the evaluation of semantic and pragmatic phenomena in order to enlarge its application domain to the dialogue evaluation.

7. References

ANDREWS, A., (1985), The major functions of the noun phrase, in T. SHOPEN (editor), *Language typology and syntactic description*, Vol. 1 Cambridge university press.

ANTOINE, Jean-Yves, SIROUX, Jacques, CAELEN, Jean, VILLANEAU, Jeanne, GOULIAN, Jerome, AHFHAF, Mohammed, (2000), Obtaining predictive results with an objective evaluation of spoken dialogue systems: experiments with the DCR assessment paradigm, LREC'2000, Athens, Greece.

ANTOINE, Jean-Yves, ZEILIGER, Jérôme, CAELEN, Jean, (1998), DQR Test suites for a qualitative evaluation of spoken dialog systems: from speech understanding to dialog strategy, Proceedings of LREC'98, Granada, Spain.

BENVENISTE, C-B., (1990), *le français parlé : études grammaticales*, Editions du CNRS, Paris.

BENVENISTE, C-B., (1997), *Approches de la langue parlée en français*, Ophrys, Paris.

BLASCO-DULBECCO, M., (1999), *Les dislocations en français contemporain : étude syntaxique*, Honoré Champion, Paris.

CAELEN, J., et al, (1997), Les corpus pour l'évaluation du dialogue homme-machine, ARC B2, Journées JST-FRANCIL, Avignon.

DUBOIS, Jean, et al, (1994), Dictionnaire de linguistique et des sciences du langage, Larousse, Paris.

GADET, F., (1989), *Le français ordinaire*, Paris : Armand Colin, 1989.

GADET, F., (1992), *Le français populaire*, Paris : Armand Colin, 1992.

HOLLARD, Solange, (1997), L'organisation des connaissances dans le dialogue orienté par la tâche, Rapport technique 1-97, GEOD CLIPS-IMAG, Grenoble.

KURDI, Mohamed-Zakaria, (2001), A spoken language understanding approach which combines the parsing robustness with the interpretation deepness, to appear in the proceedings of the International Conference on Artificial Intelligence IC-AI01, Las Vegas, USA, June 25 - 28.

MINKER, W., BENNACEF, S., (1996), Compréhension et évaluation dans le domaine ATIS, Journées d'études de la parole JEP'96, Avignon, France, 417-421.

MULLER, C., (1991), *La négation en français : syntaxe, sémantique et éléments de comparaison avec les autres langues romanes*, Librairie DROZ, Genève.

PICABIA, Lélia, (1975), *Eléments de grammaire générative :* application au français, Paris : Armon Colin.

RIEGEL, M., *et al*, (1994), *Grammaire méthodique du français*, PUF, Paris.

SABAH, Gérard, (1997), Rapport final du projet DALI (Dialogue Adaptatif : Langue et Interaction), http://herakles.imag.fr/pages_html/projets/DALI.html

TESNIERE, L., (1959), *Eléments de syntaxe structurale*, Klincksiek, Paris.

WAGNER, R-L., PINCHON, J., (1991), *Grammaire du français classique et moderne*, Hachette, Paris.

ZEILIGER, Jérôme, CAELEN, Jean, ANTOINE, Jean-Yves, (1997), Vers une méthodologie d'évaluation qualitative des systèmes de compréhension et de dialogue oral homme-machine, actes JST-FRANCIL'97, Avignon, France, 437:446.

7. Annexe 7: Les systèmes d'analyse du langage oral et leurs utilisations dans les systèmes de dialogue orienté par la tâche

Le dialogue finalisé est une interaction linguistique (souvent orale) particulièrement ciblée vers la réalisation d'une tâche qui est généralement limitée. Cet aspect finalisé de ce genre de dialogues présente un ensemble de contraintes (avantages et inconvénients) pour le choix à la fois des connaissances et des stratégies de leur traitement. Dans ce chapitre, nous allons commencer par la présentation des principales composantes d'un système de dialogue oral orienté par la tâche et ensuite nous allons passer à la présentation de plusieurs systèmes que nous avons classés selon deux étapes historiques. Cette présentation sera faite avec une mise en relief des modules d'analyse linguistique et leur interaction avec le reste des modules des systèmes dans lesquels ils sont utilisés.

7.1 Schéma général des systèmes de dialogue orientés par la tâche

Dialoguer en langue naturelle exige un système capable d'assurer un ensemble relativement considérable de fonctionnalités et de processus qui ne sont pas sans analogie avec certains processus cognitifs et moteurs humains de traitement de l'information (Calliope, 1989), (Rastier, 1990).

Etant donné l'extrême variabilité des différentes sources de connaissances, il semble que leur intégration dans un seul algorithme ou même dans un ensemble d'algorithmes correspondant chacun à une seule source de connaissance est loin d'être abordable, d'où le recours à des systèmes fortement modulaires, c'est-à-dire des systèmes dont chacun des modules correspond à son tour à un sous système modulaire s'occupant d'une source de connaissance particulière.

Pour la clarté de l'exposé, nous avons choisi de faire la présentation d'un système de dialogue homme-homme médiatisé par la machine dont l'architecture est très modulaire :

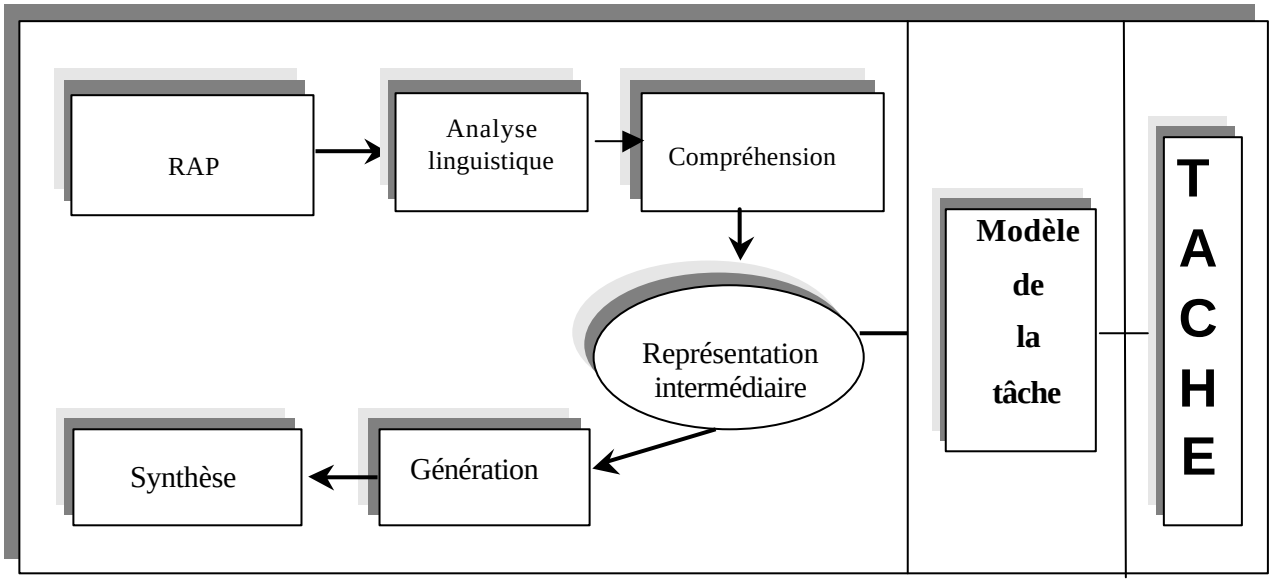


Figure 96. Architecture modulaire d'un système de dialogue homme-homme médiatisé par la machine

Afin d'être concis nous allons nous contenter de faire une présentation des différentes étapes de l'analyse et du dialogue en laissant à côté la partie qui concerne la production.

7.1.1 Reconnaissance Automatique de la Parole (RAP)

Cette phase se divise en deux parties :

4.2.3.4 Décodage acoustico-phonétique

On définit généralement le DAP comme « la mise en correspondance du signal et d'unités phonétiques prédéfinis (opération de couplage/identification) dans lequel les niveaux de représentation ressortissent du continu et du discret» (Caelen, 1991).

4.2.3.5 Modèle de langage

La fonction essentielle du modèle de langage est la restriction de la sortie du module de DAP à l'aide d'un ensemble de modèles de mots. En d'autres termes ce module a pour fonction de transformer le treillis de phonèmes fourni par le système de DAP en treillis de mots.

7.1.2 Analyse linguistique

Il s'agit ici de calculer un premier sens d'un énoncé. Ce sens n'est autre qu'une représentation logique associée à un ou à un ensemble d'énoncé(s) qui forme(nt) un seul tour de parole et dont la forme varie suivant un bon nombre de facteurs relatifs au locuteur comme l'âge, le niveau social, l'état psychologique, etc. (Rapaport, 1995).

7.1.3 Compréhension

Dans cette phase, on traite le sens élémentaire obtenu dans la phase précédente de manière plus approfondie. En général, ce traitement consiste en la situation du sens par rapport à son contexte dialogique et pragmatique notamment à l'aide de l'historique du dialogue. Le rôle de cette phase est plus important dans le contexte du dialogue homme machine que dans celui du dialogue homme-homme médiatisé par la machine étant donné que dans ce dernier cas le locuteur humain est chargé de la contextualisation des énocnés.

7.1.4 La représentation intermédiaire

Dans le contexte des systèmes de dialogue homme-homme médiatisé par la machine, la sortie du module de compréhension (ou parfois celle du module d'analyse linguistique) est une représentation intermédiaire (interlangue). Il s'agit d'un formalisme sémantique général caractérisé à la fois par son indépendance des langues et sa capacité à exprimer les différents phénomènes linguistiques qui peuvent apparaître dans n'importe quelle langue.

7.1.5 La tâche

La tâche joue plusieurs rôles dans un système de dialogue. En effet, certaines informations relatives à la tâche sont directement utilisées dans le système de dialogue alors que certaines d'autres inspirent indirectement la conception des différents modules du système. Deux Types d'informations relatives à la tâche peuvent être distingués (Pierrel, 1991) : le modèle de la tâche et l'univers de la tâche.

4.2.3.6 Le modèle de la tâche

On y distingue :

- Une définition sémantico-pragmatique des objets et des relations reliant ces objets. Ces connaissances qui sont très liées aux entités lexicales peuvent être regroupées dans un lexique spécifique à chaque application.
- Une définition en termes de buts et sous buts spécifiant les chemins d'accès aux données de l'application. Différents modèles peuvent être envisagés ici, en particulier des représentations déclaratives de type schéma (Minsky, 1975), ou réseaux sémantiques (Quillian, 1968).

4.2.3.7 L'univers de la tâche

Il s'agit d'un complément du modèle de la tâche surtout dans le cas d'un univers dynamique. En général, l'univers de la tâche doit décrire l'état de la base de donnée. Ces connaissances sont nécessaires pour déterminer la réaction du système.

7.1.6 Les problèmes des systèmes de dialogue orientés par la tâche

Les erreurs artificielles constituent avec les problèmes liés à la nature linguistique de l'entrée (que nous avons vu dans le premier chapitre de cette thèse) la principale source de problèmes au sein des systèmes de dialogue. Il s'agit, en effet, des erreurs qui sont causées par le traitement même du système de l'énoncé. Ce genre d'erreurs apparaît typiquement dans les contextes où un module prend comme entrée la sortie d'un autre module. Par exemple, une erreur d'analyse morphologique peut conduire à l'erreur l'analyseur syntaxique qui prend comme entrée cette analyse morphologique. Dans le cadre des systèmes de traitement automatique de la parole, la source principale des erreurs artificielles est la reconnaissance de la parole et ce malgré les avancées significatives dans le domaine de la reconnaissance automatique de la parole (voir (Cole, 1996) pour une revue générale de ce domaine). Trois types d'erreurs de reconnaissance de la parole sont possibles [66] :

1. **Insertion :** le système de reconnaissance insère un mot ou une série de mots qui peuvent bruiter l'analyse. L'insertion se fait généralement en ajoutant un mot dont la probabilité d'occurrence dans un contexte donné est très élevée. Prenons l'exemple suivant :

   (…) pour les nuits de lundi 9 et de mardi 10 [**Août**] (110)

   Le système a inséré le mois *août* puisque sa probabilité d'occurrence est très élevée après les mots *mardi 10*. Cela conduit l'analyseur syntaxique à une interprétation erronée. La correction de cette erreur par des règles **linguistiques** de post-traitement est impossible : l'énoncé est bien formé à tous les niveaux linguistiques : syntaxe, sémantique, pragmatique. Seules des informations issues de la tâche et de l'historique du dialogue combiné aux scores de confiance du système de reconnaissance peuvent aider dans ce cas.

2. **Suppression :** ce genre d'erreurs consiste à supprimer un mot de l'énoncé comme dans l'exemple suivant :

   (...) vers vingt trois heures (...). (énoncé de base) (111)

   (...) vers trois heures (...). (sortie de la reconnaissance) (112)

   Dans ce cas le système a supprimé le mot *vingt*, ce qui ne pose pas de problème syntaxique mais conduit à une faute d'interprétation sémantique de l'énoncé.

3. **Remplacement :** dans ce genre de cas, le système remplace un segment (un mot ou un ensemble de mots) par un autre segment généralement similaire phonétiquement comme dans l'exemple suivant.

[66] Les exemples d'erreurs de reconnaissance donnés dans cette section sont tous des cas réels du système RAPHAEL.

Au re**voir** et à bientôt	(énoncé de base)	(113)
Bon**soir** à bientôt	(sortie de la reconnaissance)	(114)

Comme nous pouvons le voir dans les deux énoncés précédents, le système de reconnaissance a remplacé la séquence *au revoir et* par *bonsoir,* ce qui donne un énoncé tout à fait correct syntaxiquement mais dont le sens est tout à fait l'inverse de l'énoncé de base (salutation ouverture/fermeture).

7.2 Présentation de quelques systèmes de dialogues orientés par la tâche

Historiquement, nous pouvons distinguer deux étapes dans le développement des systèmes de dialogue[67].

- La période des approches théoriques et expérimentales.
- Le début des applications réelles.

7.2.1 La période des approches théoriques et expérimentales

Les systèmes de cette période se distinguent globalement par les points suivants :

1. La taille du vocabulaire est très limitée (parfois moins de cinquante mots seulement). Ceci est essentiellement dû aux capacités très limitées des systèmes de l'époque (il y avait des systèmes de traitement de l'écrit dont le vocabulaire était beaucoup plus large).
2. Le lien entre l'étape d'analyse et celle de la reconnaissance était plus étroit qu'il l'est actuellement. En effet, les systèmes d'analyse linguistique du langage oral ont pour entrée non la séquence de mots mais un treillis produit par le module de décodage acoustico-phonétique qui est considéré en quelque sorte comme un module du système de compréhension au même titre que les modules d'analyse lexicale ou syntaxique par exemple.
3. Les formalismes linguistiques utilisés pour l'analyse sont peu variés. Généralement, il s'agit de grammaire indépendante du contexte, ATN ou grammaire sémantique. Ceci a changé progressivement notamment après la proposition des grammaires d'unification au début des années quatre-vingt.

[67] Cette distinction est, bien entendu, globale et approximative. En effet, d'une part les changements dans le domaine se sont fait de manière continue et progressive et d'autre part la rapidité avec laquelle ces changements ont été faits varie parfois considérablement d'un pays à l'autre.

4. Les architectures utilisées sont assez diverses : en série, hiérarchiques, hétérarchiques, tableaux noirs, etc. (voir (Erman *et al.*, 1980), (Sabah, 1989), (Carré *et al.*, 1991) pour une revue de ces techniques).
5. Au niveau des algorithmes d'analyse, les différents systèmes utilisent des approches classiques d'analyse complète (algorithmes classiques ascendants ou descendants). Ces algorithmes, malgré leur efficacité et la profondeur d'analyse qu'ils permettent, ne donnent pas suffisamment de souplesse pour le traitement des phénomènes de l'oral.

Dans ce qui suit, nous allons faire une présentation brève des principaux systèmes de cette période :

4.2.3.8 Le système MYRTILLE I

Ce système a été développé au sein du CRIN à Nancy par (Pierrel, 1975). Il est destiné à la reconnaissance de la parole continue. Pour ce faire, il met en œuvre une stratégie descendante pour le guidage du module de décodage acoustico-phonétique par les connaissances syntaxiques. Le lexique de ce système est d'environ quarante mots seulement.

4.2.3.9 Le système MYRTILLE II

Ce système est basé sur une architecture hétérarchique (Pierrel, 1978) dans laquelle il existe un module spécifique appelé contrôleur qui est responsable de la stratégie du système et qui gère les hypothèses de mots.

4.2.3.10 Le système HEARSAY II

Il s'agit d'un système de reconnaissance et analyse linguistique de la parole. Ce système est très connu dans la littérature du traitement de la parole ainsi que dans celle de l'Intelligence Artificielle. En effet, c'était le premier système qui intègre plusieurs experts spécialisés au sein d'une architecture innovante à base de tableau noir (Erman *et al.*, 1980).

4.2.3.11 Le système DIAL

Le système DIAL est un système de dialogue homme-machine basé sur une architecture multi-agents (Carbonnel et Pierrel, 1986). Les agents utilisés sont de haut niveau (des agents cognitifs). Ainsi, plusieurs agents sont utilisés selon les tâches du système comme l'agent de décodage acoustico-phonétique, l'agent de prosodie, l'agent d'analyse lexicale, l'agent d'analyse syntactico-sémantique et l'agent de dialogue.

4.2.3.12 Le système DIRA

DIRA est un système multi-experts développé au sein de l'équipe dialogue de l'Institut de la Communication Parlée *ICP* (ancêtre de l'équipe GEOD du laboratoire CLIPS-IMAG) (Nasri, 1990), (Caelen, 1990). Ce système est constitué de cinq modules correspondant à différentes sources de connaissances organisées autour d'un tableau noir supervisé.

4.2.3.13 Le système CARAMEL

Développé au sein du LIMSI, le système CARAMEL (Compréhension Automatique de Récits, Apprentissage et Modélisation des Échanges Langagiers) (Sabah, 1990) est destiné au traitement de

récits écrits mais dont la tâche est assez semblable à celle des systèmes d'analyse linguistique du langage parlé (à part les problèmes de reconnaissance). Ce système est supervisé par un *expert* utilisant des méta-règles dont les faits sont les représentations progressivement construites dans une mémoire de travail. Cela présente l'avantage de rendre l'architecture du système adaptative aux besoins du traitement. Ainsi, pour le traitement de cas simples, ne nécessitant pas d'interactions importantes, le déroulement du traitement est équivalent à une architecture en série. Dans des situations complexes, le système est comparable à des architectures hiérarchiques ou hétérarchiques. Une nouvelle version de ce système a été proposée CARAMEL-2 (Conscience, Automatismes, Réflexivité et Apprentissage pour un Modèle de l'Esprit et du Langage). Cette nouvelle version se fonde non seulement sur des contraintes informatiques mais aussi sur leurs relations avec les mécanismes cognitifs neuronaux ainsi que leurs rapports avec la conscience.

7.2.2 La période des applications réelles

L'analyse automatique de la parole spontanée a connu dans la décennie passée un gain d'intérêt considérable notamment dans le domaine des applications comme en témoigne le nombre important de projets de recherche menés notamment en Amérique du Nord, en Europe et en Asie de l'est. Plusieurs projets d'envergure ont vu le jour comme, le DARPA communicator aux Etats Unis, Verbmobil en Allemagne, les projets internationaux C-STAR I et II, les projets européens Nespole! et VICO pour n'en citer que quelques uns.

Dans cette période, les propriétés clés des systèmes peuvent être résumées dans les points suivants :

1. Une distinction plus nette des systèmes d'analyse linguistique et ceux de la reconnaissance. En effet, la généralisation des approches stochastiques de modèles de langages a rendu l'articulation des modules d'analyse linguistique et de reconnaissance beaucoup plus flexible. Par conséquent, les systèmes d'analyse linguistique et de reconnaissance sont devenus plus indépendants les uns par rapports aux autres (même s'il existe toujours des tentatives dont nous avons vu certaines dans les chapitres précédents de changer ce type d'interface en intégrant plus de syntaxe dans les modèles de langages). En effet, trois formes standards d'interaction sont utilisées dans les différentes applications :
    i- Le graphe de mots : ce type d'interface consiste à analyser directement le graphe (treillis) de mots fournit par le module de reconnaissance (voir par exemple (Staab, 1994), (Lavie, 1997)). Etant donné qu'il permet de prendre en considération toutes les variations proposées par le système de reconnaissance, cette approche permet le meilleur couplage

entre le module d'analyse linguistique et celui de reconnaissance. Cependant, à cause de la richesse des informations considérées dans l'analyse, le coût computationnel de cette approche est assez élevé.

ii- Les N-meilleures hypothèses de reconnaissance : cette approche consiste à analyser le graphe de mots du système de reconnaissance par un algorithme de recherche comme A* qui permet de trouver les meilleurs chemins dans le graphe (les meilleures hypothèses). Ainsi, chacune des hypothèses retenues (dont le nombre varie selon les choix du concepteur du système) est analysée séparément par le module d'analyse linguistique. Finalement un module d'arbitrage sélectionne la meilleure analyse selon les critères du module de reconnaissance et ceux du module d'analyse linguistique (voir par exemple les approches décrites dans (Zechner et Waibel, 1998) et dans (Kurdi, 1999)).

iii- La meilleure hypothèse de reconnaissance : il s'agit d'une version simplifiée de l'approche précédente où l'on retient uniquement une seule hypothèse.

2. Domination des architectures sérielles. Les systèmes avec des architectures multi-agents ont généralement pour but de valider des hypothèses théoriques sur la cognition plus que la création de systèmes applicatifs. Cependant, certaines exceptions peuvent être notées dans des applications de taille très importante comme celle du système Verbmobil que nous allons voir plus loin.

3. Diversité des formalismes linguistiques utilisés dans le traitement. En effet, plusieurs formalismes ont été utilisés pour l'analyse linguistique du langage oral comme LFG[68] (Antoine, 1994), GB[69] (Boufaden, 1998), LTAG (Lopez, 1999a) et (Halber, 1999), TFG (Roussel, 1999), HPSG[70] (Bonnema *et al.*, 1999) et (Uszkoreit *et al.*, 2000), Grammaire Sémantique (Minker *et al.*, 1996) et (Gavaldà, 2000).

4. Les systèmes d'analyse linguistique ont connu un développement considérable qui consiste en le passage des dialogues mono-tâches dans la première partie de la décennie précédente aux dialogues multi-tâches à partir de la seconde moitie de cette décennie. Ce changement dramatique et rapide témoigne de l'intérêt croissant accordé par les communautés académiques et industrielles au domaine. Cela implique une augmentation de la taille du lexique (le nombre du lexique est multiplié *n* fois où *n* est le nombre des tâches et des sous-tâches du dialogue) ainsi que l'espace conceptuel de la tâche du dialogue.

[68] Lexical Functional Grammar.

[69] Government and Binding theory.

[70] Head Driven Phrase Structure Grammar.

5. Création de différentes méthodes d'analyse superficielles (que nous avons vu dans le chapitre précédent) qui sont devenues très populaires dans le domaine.

Par ailleurs, nous pouvons noter que les explorations théoriques de nouvelles approches ne se sont pas arrêtées dans cette période. Nous pouvons citer à titre d'exemple le système Micro du Laboratoire CLIPS-IMAG qui est basé sur une architecture multi-agents inspirée du traitement de l'information dans le cerveau humain (Antoine, 1994), (Caillaud, 1996).

Nous allons présenter, dans les paragraphes suivants, différents systèmes que nous avons jugés à la fois représentatifs de la littérature et pertinents par rapport à notre approche. L'objectif de cette présentation étant de donner une idée générale des techniques utilisées dans l'analyse linguistique du langage oral ainsi que la situation de ces techniques dans le contexte d'application larges de dialogues orientés par la tâche.

4.2.3.14 Le projet ATIS

Le Projet ATIS (Air Travel Information Services) est l'un des projets les plus importants dans la première partie de la décennie précédente. En effet, ce projet financé par la DARPA (Advanced Research Project Agency) a mis en compétition plusieurs laboratoires nord américains de haut niveau comme AT&T labs., BBN, LCS-MIT, SRI Internationl, McGill University et ISL-CMU. Plus tard, le laboratoire LIMSI s'est joint à ce projet (Minker, 1995) et (Minker et Bennacef, 1996). La tâche des systèmes ATIS consiste à permettre aux utilisateurs d'accéder, dans des conditions de simulation, à des informations sur les vols assurés par les compagnies américaines et canadiennes.

Le schéma général des systèmes ATIS peut être représenté comme suit (Minker, 1995) :

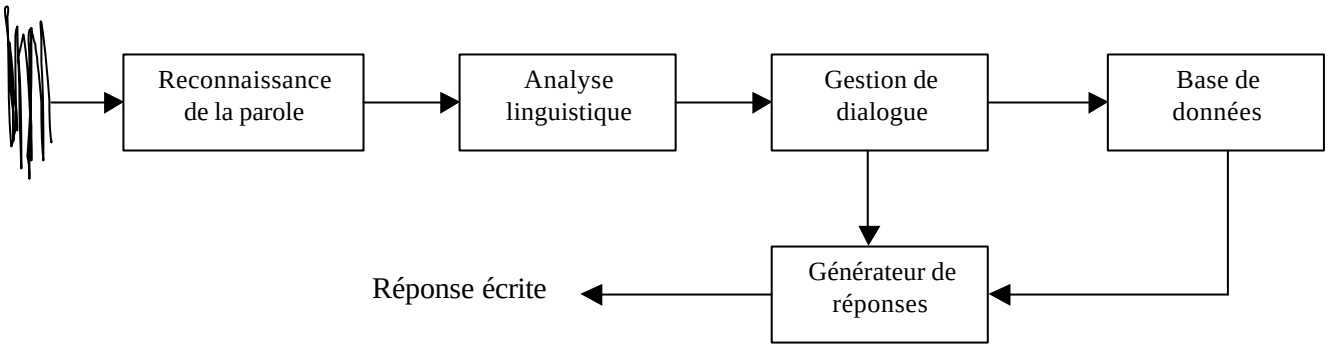


Figure 97. Le schéma général des systèmes ATIS

Dans les paragraphes suivants nous allons présenter quelques systèmes qui ont été proposés dans le cadre de ce projet avec une concentration sur les modules d'analyse linguistique étant donné que le reste des modules de ces systèmes (le gestionnaire de dialogue et le générateur) ainsi que leur

interaction avec le module d'analyse linguistique ne représentent pas d'originalités particulières. Les systèmes choisis sont basés sur des implantations variées de la grammaire sémantique.

7.2.2.1.1 Le système ATIS de AT&T

Le module d'analyse linguistique de ce système est basé sur la méthode de décodage acoustique du signal de parole. Il s'agit d'une méthode qui a été proposée par (Pieraccini et Levin, 1991) et (Pieraccini et Levin, 1995). L'originalité de cette méthode est due au fait qu'elle permet de détecter les réalisations des concepts au niveau du module de reconnaissance de la parole directement à partir du signal. Pour ce faire, la séquence d'entrée est analysée en deux phases :

1. **La phase de décodage conceptuel :** dans cette phase, afin d'associer à la chaîne d'entrée le (ou l'ensemble des) concept(s) lui correspondant, on a recours à des techniques d'analyse statistique. Soit un énoncé représenté par une séquence d'observations acoustiques : $A = a_1, a_2, \ldots a_n$. Cet énoncé correspond à une séquence de mots $M = m1, m2, \ldots Mw$. De même, chaque séquence de mots peut être associée à un (ensemble de) concept(s). Pour simplifier, le système, ayant une entrée *A*, doit trouver *M* et *C*. Ce problème peut être approché en utilisant le critère du maximum *a posteriori*, qui permet d'obtenir la probabilité conditionnelle maximale de *M* et *C* sachant *A* :

   $P(M, C|A) = \max_{MxC} P(M, C|A)$; avec la formule de Bayez :

   $P(W, C|A) = P(A|M, C)\ P(M|C)\ P(C)/P(A)$

Avec	P(A|M, C)	:	le modèle acoustique des mots
	P(W|C)	:	le modèle de langage conditionné par le concept
	P(C)	:	le modèle conceptuel

   Cette équation peut être approchée (entre autres) par les chaînes de Markov cachées HMM.

   (Pieraccini et Levin, 1995) ont utilisé dans leur système 47 états et ont fait l'entraînement sur un corpus de 547 énoncés pré-segmentés manuellement. L'apprentissage des HMMs a été réalisé par quelques itérations de l'algorithme de Viterbi.

2. **La phase de génération de traits :** dans cette phase, on traduit la représentation conceptuelle d'un énoncé en série de couples (attributs, valeurs).

   Exemple : je voudrais une chambre avec bain.

   Concept demande_information_réservation : *je voudrais une chambre*.

   Attribut caractère_de_chambre : *avec bain*.

   Le lien entre les concepts et les attributs est explicité dans une table particulière.

Cette méthode a connu plusieurs modifications et adaptations tant par ses propres auteurs dans le cadre du système de dialogue AMICA (Pieraccini et Levin, 1997) ou par d'autres chercheurs comme (Qiguang *et al.*, 1997).

7.2.2.1.2 Le système ATIS de McGill University

Le module d'analyse linguistique dans ce système est basé sur l'utilisation d'arbres de décision pour l'apprentissage de segments conceptuels d'une grammaire sémantique (appelée par l'auteur : sémantique globale) (De Mori, 1994). La fonction essentielle du module ainsi crée est la reconnaissance d'îlots de mots au sein de l'énoncé reçu. Le choix de ces îlots a été fait selon trois critères (fonctionnels) :

1. Ce sont de potentiels remplisseurs d'endroits clés qui doivent être analysés. Par exemple, ils peuvent donner l'heure, la date, etc. qui va remplir l'endroit clé dans la représentation conceptuelle.
2. Ils aident à indiquer la catégorie sémantique de l'énoncé comme la distinction entre la question et la commande.
3. Ils indiquent à quel endroit tel remplisseur appartient. Cela permet de distinguer entre l'heure de départ et l'heure de l'arrivée par exemple.

Le système fonctionne selon deux niveaux :

a. **Un analyseur syntaxique local :** cet analyseur sert à identifier les îlots de mots constituants des remplisseurs potentiels. Les deux autres types d'îlots de mots, importants sémantiquement sont difficiles à analyser à l'aide de ce module pour plusieurs raisons dont la principale est l'existence des indices (souvent lexicaux) dans des parties discontinues de l'énoncé.

b. **Des arbres de classification des chaînes :** ces arbres ont été conçus pour identifier automatiquement les îlots de mots comportant de l'information sur le format approprié pour un énoncé ou encore sur l'endroit clé approprié pour un remplisseur reconnu précédemment.

L'implémentation a été faite à l'aide d'arbres de classification binaires, qui sont des arbres dont chacun des nœuds est associé à une question oui-non, un sous arbre *OUI* et un sous arbre *NON*, et dont chacune des feuilles de nœuds correspond à une catégorie. L'algorithme d'implémentation est basé sur les éléments suivants :

- L'ensemble des questions oui-non possibles et qui peuvent être appliquées aux donnés.
- Une règle pour sélectionner la meilleure question à chaque nœud sur la base des données étudiées.
- Une méthode pour tailler les arbres afin d'éviter le sur-apprentissage.

En l'appliquant au corpus de données prélassées, cet algorithme va générer un arbre capable de classer les nouvelles données c'est-à-dire, associer à chacun des concepts l'ensemble de ses réalisations avec leurs valeurs, de manière analogue à ce qu'on a vu dans le système précédent.

4.2.3.15 Le projet DARPA Communicator

Tout comme le projet ATIS, le projet *DARPA Communicator* met en compétition les principaux laboratoires américains qui travaillent dans le domaine du traitement automatique de la parole comme le MITRE, NIST, ISL-CMU, CSLR, Colorado, IBM, ainsi que certains laboratoires européens impliqués dans des activités de recherche similaires comme le NISLab. à l'université d'Odense au Danemark[71]. Les défis ainsi que les techniques de ce projet constituent le prolongement de ceux que nous avons vu dans le projet ATIS. Afin de donner une idée sur cet avancement sans trop répéter ce que nous avons dit sur le projet ATIS, nous avons jugé bon de présenter un seul système réalisé dans le cadre de ce projet qui est le CU Communicator.

7.2.2.1.3 Le CU Communicator

Il s'agit d'un système de dialogue homme machine développé au sein du *Center for spoken Language Research* à l'université du Colorado aux Etats Unis (Pellom *et al.*, 2000) et (Pellom *et al.*, 2001).

Le système est conçu pour le traitement d'appels téléphoniques à propos d'informations touristiques : billetterie d'avions, informations sur des hôtels, réservations de voitures, etc.

Le système est composé de plusieurs modules correspondants à différents niveaux d'analyse. Ces modules sont organisés autour d'une unité centrale *hub* qui assure le lien entre les différentes composantes. Voici une présentation générale de l'architecture du système :

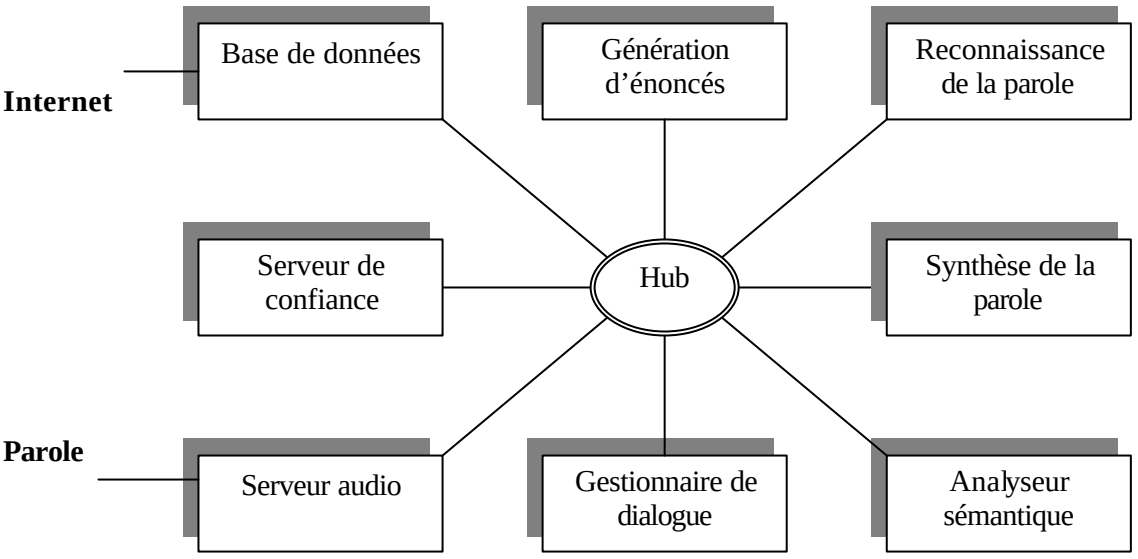


Figure 98. Architecture du CU-Communicator

[71] Voir le site de ce projet pour plus de détails sur ses différents aspects : http://www.darpa.mil/ito/research/com/

- **Le hub**

Cette unité a un rôle purement communicatif entre les modules. En effet, sa fonction se limite à transférer les informations et les données d'un module à l'autre sans effectuer un traitement quelconque sur ces données. Les messages reçus et émis par le *hub* sont codés dans un format de schémas contenant des unités élémentaires de prédicat-argument.

- **Le serveur audio**

Il s'agit du serveur qui reçoit et enregistre les messages téléphoniques. Ce système utilise le serveur DARPA développé conjointement par les MIT et le MITRE.

- **Le module de reconnaissance de la parole**

Le CU-Communicator utilise le module de reconnaissance Sphinx-II développé à Carnegie Mellon University. Il s'agit d'un reconnaisseur à base de modèle markovien semi-continu avec un modèle de langage à base de trigrammes. Une attention particulière a été accordée au traitement des noms propres, à la fois, centraux pour l'application et classiquement considérés comme une source d'erreurs de reconnaissance de la parole. Ainsi, tous les mots propres ont été étiquetés en classes comme : ville, pays, nom_aéroport, etc. Le module de reconnaissance reçoit le signal de parole du serveur audio et produit la meilleure hypothèse de reconnaissance.

- **Le serveur de confiance**

Ce module a pour fonction de filtrer les erreurs de reconnaissance au niveau des mots ainsi que les segments non pertinents pour l'application. Dans une version récente du système, ce filtre est couplé avec un autre basé sur les concepts de l'application. Ainsi, le score de confiance final est la combinaison de ceux calculés au niveau des mots et au niveau des concepts.

- **Le module d'analyse sémantique**

Une version modifiée du module d'analyse phœnix (dans le système ATIS de l'ISL-CMU) développé par W. Ward est utilisée dans le CU-Communicator. La fonction principale de ce module est d'associer à la sortie du système de reconnaissance un schéma sémantique qui contient une série de slots correspondant à des unités sémantiques pertinentes pour l'application. La différence principale de ce module par rapport à la version originale, est que la grammaire sémantique dans ce système a été aménagée de manière à permettre la représentation des différentes tâches du système de dialogue (dans le système ATIS il y a une seule tâche).

- **Le gestionnaire de dialogue**

Le gestionnaire de dialogue contrôle l'interaction entre l'utilisateur et le serveur de l'application. Ce module a une multitude de fonction dont les principales sont :

- La décision, à chaque étape de l'interaction, des actions que le système doit prendre.
- La résolution des ambiguïtés contextuelles des analyses données par phœnix. Ceci peut être fait par le lancement de requêtes de clarifications de l'utilisateur.
- Estimation du score de confiance dans les informations extraites des schémas.
- Construction de requêtes SQL.
- Envoi d'informations au module de génération.

Le gestionnaire est basé sur un modèle événementiel dans lequel le contexte dialogique joue un rôle dans la décision des actions futures.

- **La base de données et l'interface à Internet**

Il s'agit d'une base SQL et d'un ensemble de scripts orientés par le domaine pour l'accès à l'information à travers Internet.

- **Le module de génération**

Le système utilise un modèle à base de schèmes *templates* basés sur des actes de dialogue pour générer les énoncés.

- **Le module de synthèse**

Un synthétiseur concaténatif dépendant du domaine a été utilisé pour la sortie audio du système.

4.2.3.16 Le projet Verbmobil

Le projet allemand Verbmobil est, à notre connaissance, le plus grand projet jamais réalisé en Europe sur le traitement de la parole et du dialogue et l'un des plus grands projets au niveau mondial. Comme l'indique son nom, ce projet est destiné à traiter des conversations parlées à travers un téléphone mobile (Wahlster ed., 2000), (Wahlster, 2000). L'objectif principal du système construit au cours de ce projet est la traduction automatique de la parole mais des fonctions de résumés automatiques de dialogue lui ont été ajoutées. Ces fonctions permettent de générer un rappel général de ce que les deux interlocuteurs ont dit au cours de leur négociation et donc d'éviter toutes ambiguïtés ou problèmes qui peuvent résulter d'une erreur de traduction.

Le système a été réalisé dans deux phases :

1. **La première phase :** elle s'étend entre 1993 et 1996. Les propriétés clés du prototype réalisé dans cette étape sont les suivantes :
    - Le vocabulaire du système dans cette étape était d'environ 2500 mots pour la traduction allemand-anglais.
    - Un système de reconnaissance indépendant du locuteur.
    - Utilisation de la prosodie pour la désambiguïsation seulement.

- Combinaison de l'analyse superficielle et profonde pour le traitement de la parole spontanée en allemand.
- Une stratégie de clarification entre l'utilisateur et Verbmobil.
- Prise en considération du contexte dialogique dans le traitement des énoncés (contrairement à la plupart des autres approches dans le domaine de la traduction de la parole).
- L'évaluation de ce système sur vingt-cinq mille cas de traduction a montré que 74,2% des traductions produites par ce système sont *approximativement* correctes. C'est-à-dire que, dans ces cas, le sens global de l'énoncé de départ a été exprimé dans l'énoncé généré dans la langue cible[72].

2. **La deuxième phase :** cette phase impliquant trente et un partenaires industriels et académiques et environ mille deux cent chercheurs, cette phase s'est étendue entre 1997 et 2000 (Karger et Wahlster, 2000). Les principaux objectifs de cette étape sont les suivants :
   - **Multifonctionnalité :** le système doit être facilement adaptable à de nouveaux domaines de discours.
   - **Multilinguisme :** le système doit traduire des dialogues dans différentes langues. En effet, la version finale du projet traduit des textes entre trois langues de manière bidirectionnelle : Allemand-Anglais-Allemand et Allemand-Japonais-Allemand.
   - **Multimodalité :** le système doit offrir une aide à la traduction dans un contexte d'application multimodales.
   - **Mobilité :** le système doit être capable de faire des traductions à travers un serveur accessible par les téléphones mobiles.
   - **Traitement des conversations multilatérales :** le système doit être capable de traiter des dialogues entre plus de deux personnes.

   Ainsi, le traitement de ces dialogues se fait de manière entièrement centralisée dans le serveur principal du système. Comme nous pouvons le voir dans la figure 34, le premier locuteur émet son énoncé via son téléphone portable. Le signal émis est ensuite transmis par le stellite au serveur central qui effectue la traduction automatique et produit un énoncé synthétisé dans la

[72] voir : http://verbmobil.dfki.de/verbmobil/VM2.info.us.html

langue cible correspondant à l'énoncé reçu. Finalement, l'énoncé synthétisé est transmis via le satellite au mobile du destinataire. Cela veut dire que la traduction ne nécessite, à part le téléphone portable, aucun PC ou autre outil informatique.

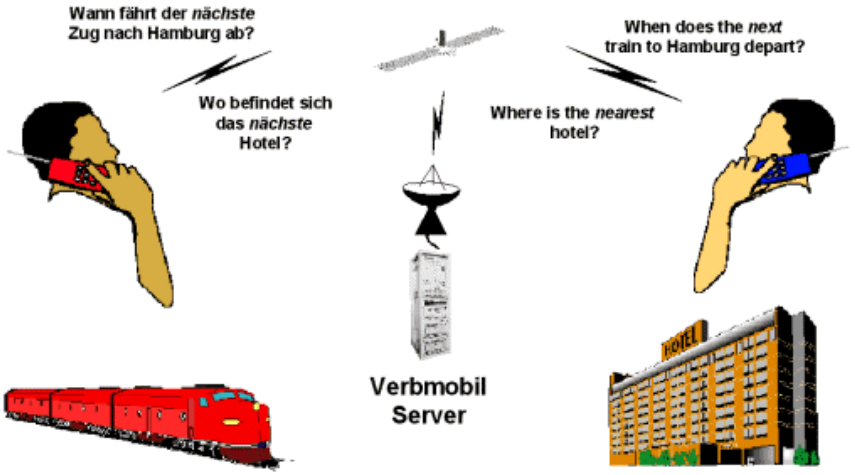


Figure 99. Schéma d'un dialogue médiatisé dans Verbmobil (Wahlster, 2000)

Verbmobil utilise 69 modules destinés à des traitements de natures assez diverses. L'une des spécificités de ce système est la parallélisation de différents modules qui ont la même fonction et la fusion de leurs résultats afin de combiner leurs avantages et de réduire leurs inconvénients. Dans la figure 35, nous présentons l'interface du système final qui donne une idée des principales composantes du système ainsi que de leurs interactions.

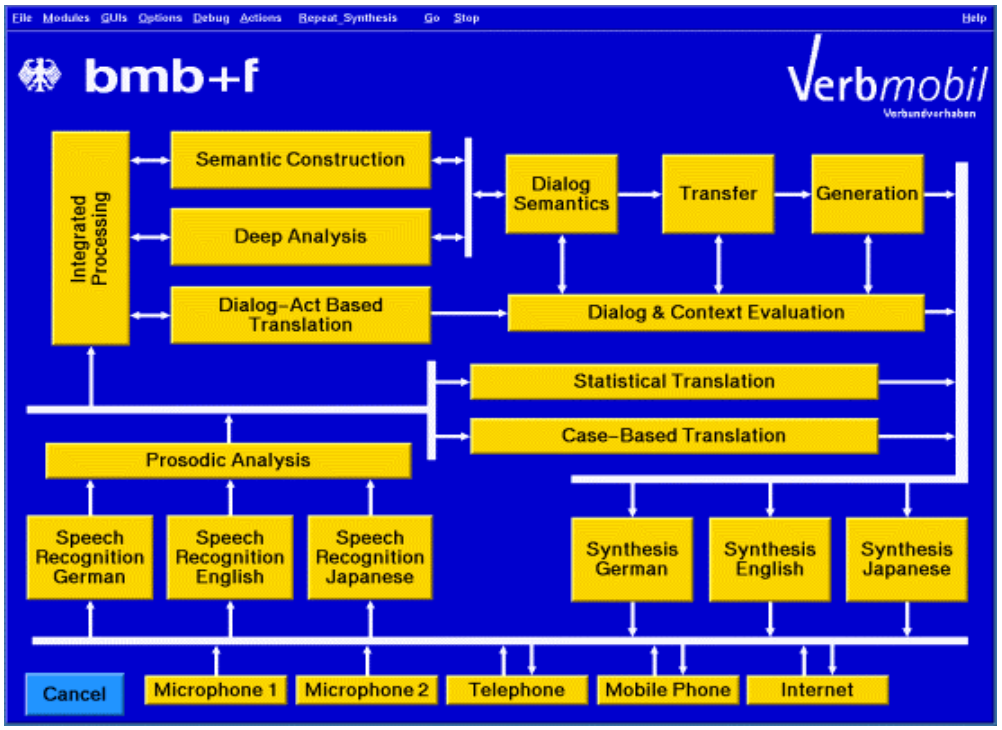


Figure 100. L'interface du système Verbmobil (Wahlster, 2000)

Dans ce qui suit, nous allons présenter les principales composantes de Verbmobil qui sont pertinentes par rapport à notre tâche.

7.2.2.1.4 L'architecture de Verbmobil

La figure précédente présente l'interface du système ainsi que ses principales fonctionnalités mais elle ne donne pas son architecture réelle qui est beaucoup plus complexe. En effet, la version finale de Verbmobil contient soixante-neuf modules qui interagissent chacun avec au moins un autre module. Etant donné l'approche de traitements parallèles et les contraintes de traitement en temps réel, les besoins de communication entre les différents modules sont énormes d'un point de vue quantité. Par ailleurs, vu l'aspect non-séquentiel de Verbmobil, cela implique que les modules échangent non seulement les entrées et les sorties de chaque module mais aussi les attentes de haut niveau top-down, des contraintes, des alternatives, des scores de confiance, des probabilités, etc.

Pour assurer œs besoins, une architecture multi-tableaux noirs a été implantée. Cette architecture contient trois éléments de types différents de composantes (Klüter, et *al.*, 2000) :

3. Un ensemble de modules indépendants appelés sources de connaissance. Ces modules sont l'élément principal de la résolution du problème.

4. Les tableaux noirs. En effet, contrairement à la plupart des architectures à base de tableau noir, Verbmobil utilise une série de tableaux utilisés chacun pour représenter les résultats intermédiaires à chaque étape du traitement. Cent-quatre-vingt-seize tableaux noirs sont utilisés pour assurer l'interaction entre les différents modules.
5. Un module de contrôle qui a pour fonction de faire des décisions sur l'allocation des sources pour optimiser le temps de calcul ainsi que le choix de l'ordre selon lequel les modules doivent intervenir.

Les propriétés clés de cette architecture sont les suivantes :

- En général, chaque module communique avec plus d'un tableau noir.
- Un module ne peut pas communiquer avec un autre module directement.
- Les modules peuvent être multiplier selon les besoins. Par exemple, deux modules de reconnaissances pour l'allemand sont utilisés pour le traitement des conversations multilatérales impliquant deux allemands.

Malgré sa complexité, cette architecture s'est révélée plus efficace et plus adaptée qu'une architecture multi-agents (plus simple) qui a été utilisée dans la première phase du projet.

7.2.2.1.5 La reconnaissance automatique de la parole

Trois systèmes de reconnaissance sont utilisés pour les trois langues de l'application. Chacun de ces systèmes, est conçu pour traiter des données avec deux degrés d'échantillonnage de transmission des données via le réseau GSM : 8 kHz et 16 kHz. La sortie du système de reconnaissance est un graphe de mots probabilisé.

7.2.2.1.6 Traitement prosodique

Selon ses concepteurs, Verbmobil est le premier projet qui utilise la prosodie systématiquement dans toutes les étapes de l'analyse (Wahlster, 2000).

Le module de traitement prosodique a pour entrée à la fois le signal de parole ainsi que le graphe de mots produit par le système de reconnaissance pour le même signal de parole. La sortie de ce module est un graphe de mots enrichi par des annotations prosodiques. L'annotation se fait sur les unités supra-phonémiques comme la syllabe, le mot, le syntagme ou le tour de parole tout entier (Batliner, 2000). La nature des étiquettes associées aux différentes unités varie selon les besoins. Par exemple, certaines unités sont annotées par la durée, le pitch, le rythme, le débit de la parole, la qualité de la voix, les pauses, etc.

Ainsi, les résultats du module multilingue de traitement de la prosodie sont utilisés pour l'analyse syntaxique (détection des frontières des unités syntaxiques, classification des énoncés selon leur mode), le traitement des dialogues (détection des actes de dialogue), la traduction, la génération et la synthèse de la parole. Cela contribue à augmenter considérablement la qualité de la traduction produite par le système dans la mesure où l'on prend en considération toutes les variations prosodiques qui ont

un effet sur le sens de l'énoncé à traduire ainsi que l'énoncé produit par le système dans la langue cible.

7.2.2.1.7 L'approche multi-moteur pour l'analyse syntaxique robuste

Trois modules d'analyse syntaxique ont été utilisés dans le traitement :

1. **Un analyseur LR stochastique :** il s'agit d'un analyseur LR (Left to Right) à la fois stochastique et incrémental pour le traitement de l'oral (Ruland, 2000). Cet analyseur est inspiré, entre autres, des travaux de Ted Brisoe à l'université de Cambridge sur la probabilisation d'analyseurs de type LR et leur extension pour le traitement des grammaires à base d'unification. Le choix de l'algorithme LR est essentiellement motivé par son efficacité et son adaptation au traitement du treillis des mots dans le graphe fourni par le système de reconnaissance.

   Les propriétés clés de cet algorithme sont les suivantes :

   - Extension du modèle probabiliste de l'analyseur en y intégrant des connaissances contextuelles probabilisées. Cet aspect a été principalement influencé par les travaux de Rens Bod sur le modèle DOP de (Bod, 1995).
   - Amélioration de la qualité de l'analyse en utilisant une phase de post-traitement des règles de transformation d'arbres. Ces règles sont apprises automatiquement à partir de corpus avec la méthode d'apprentissage par transformation *tnasformation-based learning* utilisée pour la première fois dans le domaine du traitement automatique des langues naturelles par Eric Brill pour la construction de systèmes d'analyse morphologique (Brill, 1993).

2. **Un analyseur partiel par segments (chunker) :** basé sur le système CASS de (Abney, 1991), (Abney, 1995), cet analyseur utilise des techniques d'apprentissage à base de mémoire. Ses propriétés clés sont les suivantes (Hinrichs *et al.*, 2000) :

   - Stratégie d'analyse incrémentale afin de satisfaire les contraintes imposées par Verbmobil.
   - Une attention particulière a été accorée à l'assemblage des segments, sujet pas très abordé dans le cadre des approches d'analyse partielle. Ceci permet de faciliter le plus possible la tâche du module d'analyse sémantique et par conséquent obtenir une analyse de qualité meilleure.

- Constructions de larges grammaires à états-finis pour l'allemand et l'anglais afin d'assurer une bonne couverture des phénomènes linguistiques dans la tâche de Verbmobil.

Comparé aux deux autres analyseurs, cette approche donne les meilleurs résultats d'analyse en terme de robustesse mais l'analyse qu'elle fournit est la moins profonde.

3. **Un analyseur syntaxique profond basé sur le formalisme HPSG :** cet analyseur a été réalisé dans le cadre d'une collaboration entre le DFKI à Saarbrücken, le CSLI à Stanford et le *Language Processing Lab*. à l'université de Tokyo (Uszkoreit *et al.*, 2000). L'un des principaux problèmes qui ont affronté cet analyseur est la réduction du temps de calcul nécessaire. En effet, les traits utilisés pour représenter les contraintes linguistiques dans le cadre du formalisme HPSG nécessitent beaucoup de calculs qui rendent le temps d'exécution de l'analyseur trop lent pour être intégré dans le cadre dans une application en temps réel. Deux solutions ont été combinées pour résoudre ce problème : d'une part l'élimination des traits non centraux dans le traitement ainsi que les traits disjonctifs (qui augmentent la complexité de traitement) et d'autre part le test de plusieurs types d'implantation comme, entre autres, des approximations en automates à états finis des grammaires. Finalement, ce problème a été surmonté par les chercheurs du département de Linguistique Computationnelle à l'université de Saarbrücken qui ont combiné plusieurs solutions proposées par les différents participants et ont réussi à satisfaire les contraintes du temps d'analyse (0.45 secondes pour un énoncé de 10 mots). Comme on s'attendait, les résultats de l'évaluation ont montré que ce système, comparé aux deux autres, fournit l'analyse la plus profonde mais la moins robuste aux problèmes de reconnaissance ou aux différentes extragrammaticalités de l'oral (voir (Müller et Kasper, 2000), (Flickinger *et al.*, 2000) et (Siegel, 2000) pour les résultats finaux de cet analyseur sur l'allemand, l'anglais et le japonais respectivement).

4. **Interaction des trois analyseurs :** les trois analyseurs traitent le même graphe de mots enrichis par des annotations prosodiques produites par le système de reconnaissance. Les trois analyseurs sont aussi guidés par un algorithme de type A* pour le choix des chemins les plus probables d'un point de vue reconnaissance de la parole (Ulrich et Ruland, 2000). Le schéma de la combinaison des trois analyseurs est présenté dans la figure 36 :

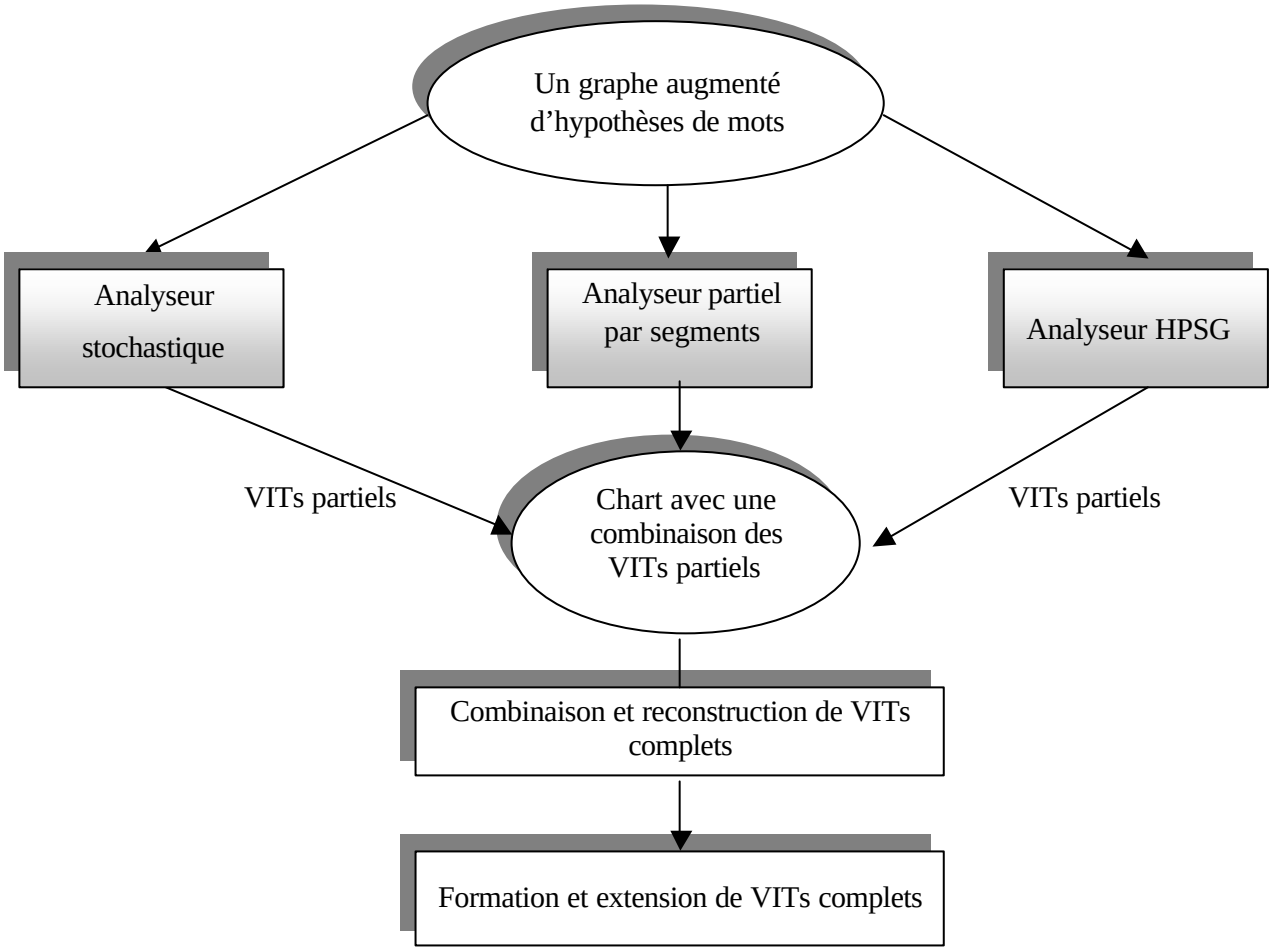


Figure 101. L'approche multi-moteur de Verbmobil

Les trois analyseurs utilisent un module d'analyse spécifique pour construire une représentation sémantique correspondant à la sortie de chacun. Ceci s'applique tant aux analyses complètes qu'aux analyses partielles. Les données échangées entre les différents modules sont représentées au format VIT *Verbmobil Interface Terms* qui est, comme son nom l'indique, un format spécial destiné à uniformiser les sorties des analyseurs. Cette conversion des sorties des analyseurs en représentation sémantique standardisée rend possible des opérations de post-traitement visant la sélection de la meilleure analyse fournie par le système (Shiehlen *et al.*, 200), (Rupp *et al.*, 2000).

Résumé Cette thèse porte sur le traitement du langage oral spontané dans le contexte du dialogue homme-machine. En partant du constat que l'usage de la langue orale s'écarte d'une "bonne" syntaxe de l'écrit, des méthodes de traitement particulières sont alors développées pour adresser des phénomènes grammaticaux et extragrammaticaux comme les répétitions, hésitations, auto-corrections, faux-départs, etc. Une approche de traitement des extragrammaticalités basée sur l'analyse d'un corpus ainsi qu'un formalisme grammatical pour l'oral (Sm-TAG) sont proposés et implémentés dans trois outils : Corrector, Oasis et Navigator. Les résultats d'évaluations quantitatives et qualitatives de ces outils sont donnés et commentés.

Mots clés Analyse linguistique, robustesse, syntaxe du langage oral, extragrammaticalités du langage oral, grammaires d'arbres et Grammaire sémantique.

Abstract Spontaneous Spoken Language SSL presents many differences compared to the written one. These differences are observed both in tem of grammatical and extragrammatical phenomena like repetitions, self-corrections, false-starts, etc. This thesis addresses the problem of parsing SSL in the context of human-machine dialogue from two points of view: theory and application. First, a corpus study of spoken language extragrammaticalities is done and a linguistic formalism (Sm-TAG) is proposed. Then the results of the theoretical work are used in the implementation of the systems Corrector, Oasis and Navigator. Evaluation of these systems following quantitative and qualitative methods is done.

Key Words Parsing, robustness, spoken language syntax, spoken language extragrammaticalities, tree grammars and semantic grammar.