Étude d'un analyseur de surface de la langue naturelle : application à l'indexation automatique de textes

Patrick Palmer

▶ **To cite this version:**

Patrick Palmer. Étude d'un analyseur de surface de la langue naturelle : application à l'indexation automatique de textes. Modélisation et simulation. Université Joseph-Fourier - Grenoble I, 1990. Français. NNT : . tel-00337917

HAL Id: tel-00337917
https://theses.hal.science/tel-00337917

Submitted on 10 Nov 2008

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

THESE

présentée par

Patrick PALMER

pour obtenir le titre de **docteur**
de l'**Université Joseph Fourier - Grenoble I**
(arrêté ministériel du 5 juillet 1984)

spécialité Informatique

Etude d'un analyseur de surface de la langue naturelle.

Application à l'indexation automatique de textes.

date de soutenance :	3 septembre 1990	
composition du jury :	président :	M. Christian BOITET
	rapporteurs :	M. Jacques COURTIN
		M. Patrice POGNAN
	examinateurs :	M. Yves CHIARAMELLA
		M. Claude DELVIGNA

Thèse préparée au sein du Laboratoire de Génie Informatique - IMAG
à l'Université Joseph Fourier - Grenoble I

SOMMAIRE

Chapitre V : L'ANALYSEUR DE SURFACE

INTRODUCTION

Ce travail a été réalisé dans le cadre du projet IOTA (développement d'un prototype de système intelligent de recherche d'informations) au sein de l'équipe SIRI (Systèmes Intelligents de Recherche d'Informations) du laboratoire de Génie Informatique de Grenoble, et a pour objet la réalisation d'un module d'analyse linguistique destiné à permettre une indexation automatique de textes en langue naturelle.

La problématique de la recherche d'informations consiste à partir de l'expression d'une requête d'un utilisateur, qui définit le thème de la recherche, à retrouver un ensemble de documents (ou de références à des documents) dont le contenu sémantique correspond le mieux au thème recherché, parmi l'ensemble des documents de la base documentaire considérée. Cette recherche s'effectue par l'exploitation d'une fonction de correspondance qui établit une association entre un thème de recherche (requête) et un contenu sémantique (documents). Cette fonction de correspondance est fondée sur la définition de modèles sémantiques pour les thèmes et les documents, et sur la définition de critères de correspondance entre ces éléments.

Le schéma ci-après, tiré de [CHIA 88], permet de représenter synthétiquement la problématique des systèmes de recherche d'informations (SRI) :

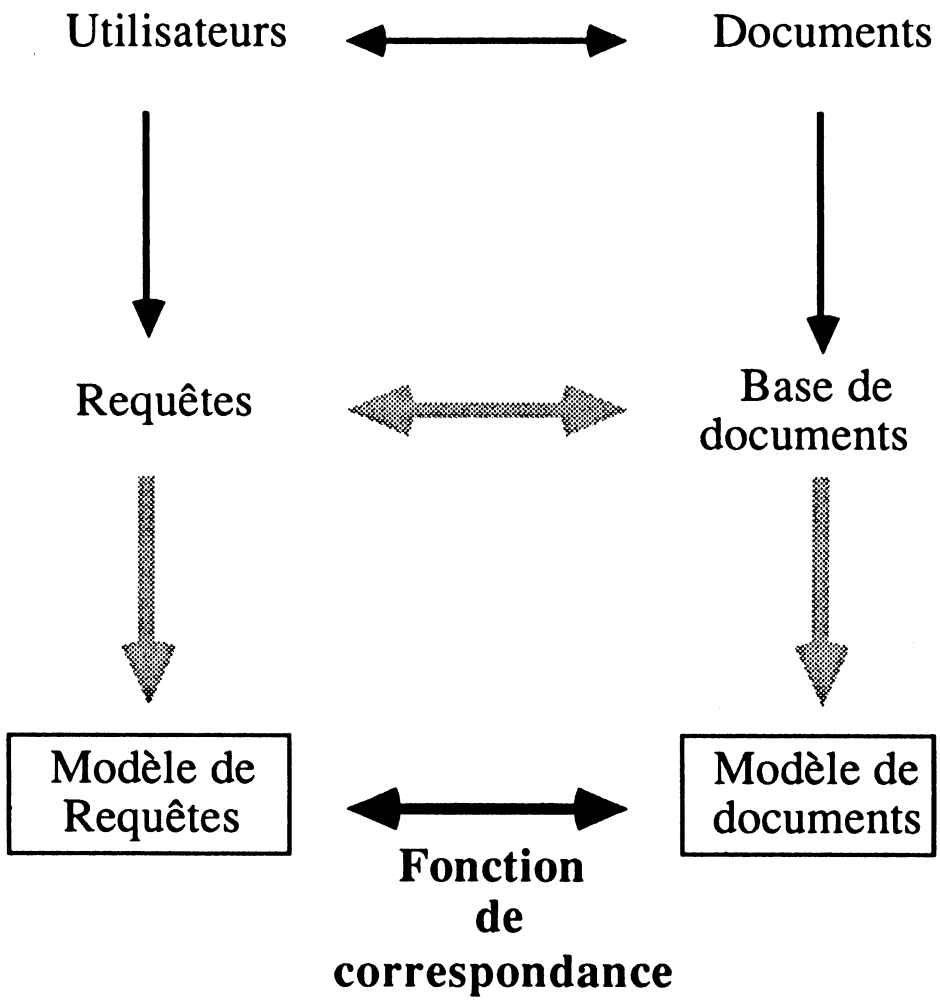


Figure 1 : *Problématique des SRI*

La représentation d'un document dans le modèle sémantique correspondant est réalisée par la fonction d'indexation du système de recherche d'informations, la compréhension de la requête et l'évaluation de la fonction de correspondance sont réalisées par la fonction d'interrogation du SRI. Nous présentons les différents composants des fonctions d'indexation et d'interrogation d'un système de recherche d'informations dans le **premier chapitre**.

L'objectif d'un module d'analyse linguistique pour la réalisation de la fonction d'indexation d'un SRI, est de permettre la représentation du contenu sémantique des documents, par la reconnaissance automatique, à partir de leur contenu textuel, de concepts structurés (mots ou groupes de mots) susceptibles de constituer des éléments du modèle sémantique.

Nous nous intéressons ici à un modèle sémantique simple dans son principe : le contenu de chaque document est représenté par un ensemble de termes (appelés termes d'indexation), qui correspondent à autant de concepts significatifs du contenu sémantique du document. Le degré d'élaboration du modèle sémantique des documents qui conditionne l'efficacité de la fonction de correspondance et donc l'efficacité du SRI, est donc directement dépendant du niveau de représentation des termes d'indexation. Ce niveau pouvant être plus ou moins proche de celui du langage naturel : mots isolés, syntagmes, phrases ...

Nous nommons Groupes Conceptuels (G.C.) la représentation normalisée, dans un formalisme cible, de ces concepts structurés qui correspondent dans IOTA aux groupes nominaux. La présentation des groupes conceptuels et des différents choix ayant conduit à leur définition est effectuée dans le **troisième chapitre**.

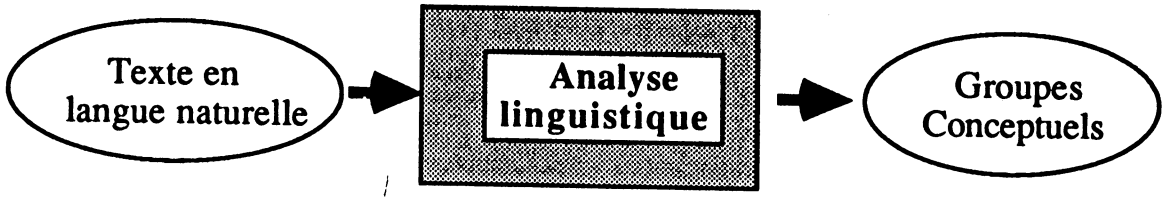


Figure 2 : *Analyse linguistique*

L'analyse linguistique mise en oeuvre dans un tel contexte (fonction d'indexation d'un SRI) doit permettre tout d'abord, l'appréhension de domaines textuels ouverts (sans limitation du vocabulaire ou des tournures stylistiques), afin que le système de recherche d'informations conserve une certaine généralité et puisse donc être utilisé pour différents types de corpus (collections de documents). Il nous parait essentiel à ce propos, de ne pas figer le vocabulaire utilisable par la donnée a priori d'informations trop fines pour pouvoir être ensuite retrouvées automatiquement (lors de la rencontre de mots nouveaux), telles que des informations de type sous-catégorisation syntaxique ou classification sémantique. C'est ce que nous appellerons le caractère général

de l'analyse. Nous aurons l'occasion d'y revenir au **deuxième chapitre** lors de la présentation des traitements automatiques du français, utilisés dans le domaine de la recherche d'informations.

Cette analyse doit également permettre le traitement de corpus de taille réelle; ce qui nécessite des algorithmes performants, et ce qui exclut a priori l'utilisation de mécanismes complexes et coûteux notamment pour l'analyse linguistique, tels que le réclament les analyses syntaxiques complètes des langues naturelles. Ce type d'analyse fournissant les différentes structures syntaxiques potentielles des phrases rencontrées, constitue dans notre contexte, un traitement disproportionné par rapport au résultat escompté, qui consiste à reconnaître et à traiter des syntagmes nominaux, pour l'extraction de groupes conceptuels.

Ces différentes considérations nous ont amené à nous orienter vers une analyse linguistique partielle des textes d'entrée, en nous limitant à la structure de surface des phrases de ces textes.

Nous présentons au cours du **quatrième chapitre** le modèle linguistique utilisé pour la réalisation de cette analyse, et nous donnons une définition formelle des données manipulées au cours des différentes étapes de ce processus. Le modèle linguistique est fondé sur la classification syntaxique habituelle de la langue française; les catégories grammaticales sont en nombre relativement restreint (une cinquantaine) et ont la particularité d'être diversifiées uniquement pour les classes fermées du français. Cette simplicité doit permettre une utilisation aisée de notre système d'analyse et une adaptabilité à une autre classification simplifiée.

Les principales caractéristiques de cette analyse linguistique de surface sont constituées par les modules suivants, qui sont détaillés au **cinquième chapitre** :

-1) Une analyse morphologique réalisant la segmentation des textes d'entrée en formes; ces formes étant composées d'une racine et d'une désinence. Ce processus utilise un dictionnaire de racines, un ensemble de désinences prédéfinies, et un ensemble de modèles morphologiques permettant de valider pour chaque solution morphologique d'une forme, la composition racine-désinence. La principale particularité de cette analyse morphologique est de s'effectuer sans retour arrière, et cela grâce à la structure particulière du dictionnaire d'analyse où les racines sont factorisées.

-2) Un processus de levée des ambiguïtés grammaticales générées par l'analyse morphologique. Nous nous sommes appliqué pour cette résolution à utiliser et développer des techniques peu coûteuses (au sens complexité algorithmique), de manière à rester cohérent avec notre objectif de reconnaissance à moindre coût. Cette résolution n'est que partielle puisque nous ne disposons ni de connaissances de type sémantique, ni des structures syntaxiques potentielles des phrases analysées. Nous procédons en deux étapes :

-a) Tout d'abord par la réalisation d'un filtrage syntaxique basé sur l'utilisation de relations positionnelles binaires, qui sont codées à l'aide d'une matrice de précédence binaire multivaluée, permettant de valider ou d'invalider la succession des solutions morphologiques de deux formes consécutives du texte analysé.

-b) Puis par l'application de schémas de résolution d'ambiguïté prédéfinis qui sont consignés dans un catalogue. Ces schémas sont constitués de deux parties : une première partie décrivant l'ambiguïté, dont la reconnaissance dans le graphe des solutions morphologiques entraîne l'application de la deuxième partie décrivant la simplification de ce graphe.

-3) Un processus d'enrichissement automatique du vocabulaire; la nécessité d'un tel processus est évidente si l'on veut pouvoir traiter de manière non restrictive divers domaines d'application. Dans notre système d'analyse, il s'agit en fait d'un traitement des formes inconnues, qui permet de déterminer automatiquement un ensemble de solutions morphologiques potentielles en fonction de l'orthographe et du contexte grammatical proche de ces formes. Lorsque cette détermination est non ambiguë, ces formes sont consignées dans le dictionnaire d'analyse avec le modèle morphologique associé. Ce processus d'enrichissement ne porte que sur les classes ouvertes du français et utilise le fait que l'ensemble des formes appartenant aux classes fermées, ou ayant un comportement grammatical irrégulier ou singulier (accord en nombre, conjugaison) parmi les classes ouvertes, a été consigné lors de l'initialisation du système (nous présentons les notions de classes ouvertes et fermées dans notre modèle linguistique au quatrième chapitre).

L'implantation de cet analyseur de surface et les résultats de notre expérimentation sur des textes issus de corpus de types différents, sont présentés au cours du **sixième chapitre**.

Notre travail aborde donc le problème de l'indexation automatique de documents dans sa partie amont, qui consiste à extraire des documents textuels

des entités susceptibles de véhiculer les concepts significatifs de leur contenu sémantique. L'apport de notre étude par rapport aux méthodes existantes réside dans un traitement efficace de la reconnaissance des groupes nominaux. Cette efficacité est recherchée sur trois points : une analyse de surface permettant de limiter les traitements, une résolution locale des ambiguïtés, des possibilités d'apprentissage de mots inconnus.

CHAPITRE I

PLAN DU CHAPITRE I

LES SYSTEMES DE RECHERCHE D'INFORMATIONS

LES SYSTEMES DE RECHERCHE D'INFORMATIONS

1. Introduction

Nous présentons dans ce chapitre ce qu'il est convenu d'appeler le domaine d'application de notre étude : les Systèmes de Recherche d'Informations, ou SRI. Cet exposé va s'effectuer en deux parties :

La première consistera à situer le domaine de la recherche d'informations, domaine pluridisciplinaire intégrant des techniques issues de différentes disciplines; aussi bien des bases de données, que de l'intelligence artificielle, de la linguistique que de la cognitique, etc... Nous commencerons par en exposer les objectifs et les fonctionnalités, avant de décrire les principales composantes des systèmes de recherche d'informations. Nous poursuivrons par un bref survol des systèmes classiques, pour lesquels nous nous sommes restreint à la présentation de deux des principaux modèles utilisés dans les composantes d'indexation automatique de textes en langue naturelle : le modèle statistique et le modèle linguistique.

Dans la deuxième partie, nous présenterons les travaux menés dans le cadre du projet IOTA, afin de mieux situer notre travail dans le contexte du

développement d'un prototype de Système intelligent de recherche d'informations. La composante d'indexation de ce prototype constitue le support de notre réalisation.

2. Problématique des SRI

Un système de recherche d'informations a pour fonction principale de permettre de retrouver (ou de localiser) dans un corpus particulier, des entités (généralement textuelles : document, section, paragraphe, ...) "*correspondant*" à des besoins exprimés par un usager.

Cette correspondance implique une fonction de comparaison entre les thèmes évoqués dans la requête (demande, question d'un utilisateur) et le contenu sémantique des documents; la définition et l'implémentation de cette fonction constituent l'essentiel de la problématique de la recherche d'informations. Elles sont fondées sur la définition d'un modèle de représentation du contenu des documents et des requêtes, et d'une fonction de comparaison non stricte introduisant la notion de pertinence d'un document par rapport à la requête.

La problématique des SRI consiste donc à identifier parmi l'ensemble des thèmes contenus dans les documents formant le corpus considéré, celui ou ceux recherchés par un usager, qui constituent ses besoins, et à fournir ensuite en sortie, comme résultat, le moyen d'y accéder (consultation, localisation ...).

Classiquement, les besoins de l'utilisateur sont exprimés au moyen de requêtes d'interrogation, le résultat étant constitué, dans le meilleur des cas, par les entités textuelles les plus pertinentes par rapport aux thèmes identifiés, ou bien, plus généralement, par une liste de références textuelles permettant à l'utilisateur de cerner sa recherche, c'est-à-dire d'accéder aux documents.

On peut représenter schématiquement le fonctionnement d'un système de recherche d'informations de la façon suivante :

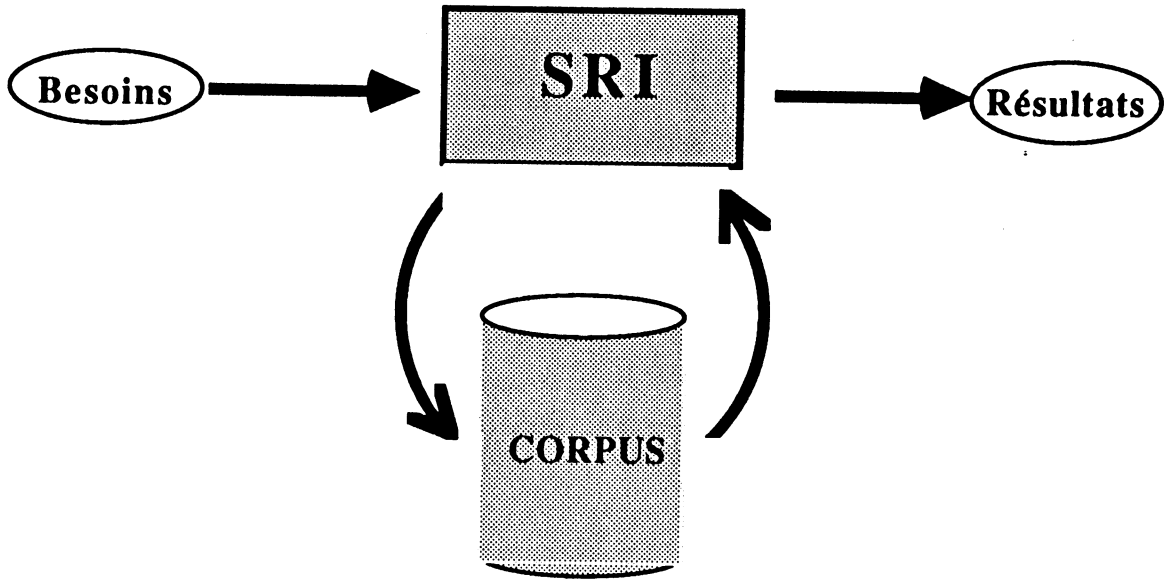


Figure 1 : *Fonctionnement d'un SRI*

Pour qu'un tel système puisse être utilisé efficacement par des usagers non-spécialistes et convenir à un grand nombre d'applications, il est nécessaire que certaines conditions assurant une bonne convivialité soient remplies. Les principales sont :

- La capacité de traiter des corpus volumineux en maintenant un temps de réponse raisonnable (les plus gros serveurs actuels manipulent plusieurs millions de références textuelles).

- Une bonne qualité des réponses, (i.e. une bonne "pertinence" moyenne des références fournies par le système).

- Un formalisme d'expression des requêtes approprié.

On peut ajouter à ces trois conditions une réelle transparence pour l'utilisateur, ce qui n'est pas toujours aisé à réaliser : la connaissance des formalismes utilisés et des niveaux de représentation internes est souvent nécessaire pour une bonne utilisation d'un système donné.

La pertinence de l'adéquation entre l'ensemble des réponses fournies en sortie par le système et les besoins des usagers exprimés par des requêtes est un facteur qualitatif essentiel des SRI. Si le système est optimal, cet ensemble doit comprendre toutes les réponses pertinentes possibles du corpus traité, et

uniquement celles-ci. Dans le jargon documentaire, le taux des réponses parasites par rapport à l'ensemble des réponses s'appelle le "bruit", et le taux des réponses pertinentes non retrouvées pour une requête, se nomme le "silence". Optimiser l'efficacité qualitative d'un SRI, revient à minimiser ces deux quantités.

3. Les composantes d'un SRI

Un SRI peut se décomposer en deux fonctions principales complémentaires, qui sont réalisées par la composante d'indexation et par la composante d'interrogation. La base de connaissances, structurée généralement en thésaurus, est étroitement liée à ces composantes, mais sa définition constitue à elle seule un aspect particulier (cf. 3.3).

La nécessité de puissance de calcul du système pour garantir un temps de réponse satisfaisant les usagers exclut les méthodes de traitement complet en temps réel du corpus, dès que celui-ci est un tant soit peu volumineux (cas les plus fréquents). En conséquence, la solution "universellement" adoptée est de pré-traiter l'ensemble des documents du corpus pour identifier dans un premier temps l'ensemble des concepts véhiculés par les documents, et de sélectionner ensuite pour chacun d'eux, ceux jugés les plus représentatifs de ce contenu (mise en oeuvre du modèle de représentation du contenu des documents). Ce pré-traitement constitue la composante d'**Indexation** du SRI, que l'on peut voir comme la succession de trois phases :

- une phase d'analyse qui permet de reconnaître et d'extraire les concepts présents dans les documents,

- une phase de normalisation et de sélection des concepts représentatifs,

- et enfin une phase de stockage, ou de construction de la relation d'indexation qui établit un lien entre les concepts retenus, appelés termes d'indexation, jugés représentatifs, et les documents concernés.

Les utilisateurs formulent leurs besoins au moyen de requêtes d'interrogation exprimées selon un formalisme qui doit être analysable et "compréhensible" par le système. Cette analyse de la requête et son interprétation constituent la composante d'**Interrogation** du SRI.

Le confort des usagers (transparence et possibilités d'expression) est étroitement dépendant de la souplesse de ce formalisme. Inversement, le processus d'analyse (ou d'interprétation des requêtes) sera d'autant plus complexe que le formalisme d'expression sera plus libre. Ce formalisme peut aller d'une liberté totale avec l'utilisation du langage naturel jusqu'à des langages d'interrogation très formels, et donc difficiles d'accès pour des usagers non spécialistes. Quel que soit le formalisme utilisé, le système doit pouvoir interpréter les requêtes d'interrogation afin de pouvoir confronter les besoins exprimés par les utilisateurs avec les concepts retenus lors de la phase d'indexation pour représenter le contenu des documents du corpus (fonction de correspondance question-document évoquée plus haut).

Afin de permettre cette confrontation, il est nécessaire de représenter dans un formalisme interne équivalent, les résultats de la phase d'indexation et de l'interprétation des requêtes. Cette représentation interne (ou modèle sémantique) peut permettre ensuite divers traitements sémantiques, afin de faciliter le calcul d'une correspondance entre les contenus, des requêtes et des documents indexés. Ces calculs nécessitent une certaine connaissance sur le domaine couvert par le corpus, qui est stockée dans un **thésaurus**, que l'on peut définir comme un lexique de concepts reliés entre eux par des relations sémantiques. La constitution d'un thésaurus est un problème très délicat, souvent résolu manuellement par des spécialistes du domaine concerné.

Le thésaurus peut être utilisé pour normaliser les termes d'indexation lors de la phase d'indexation, et les requêtes lors de la phase d'interrogation. Son utilisation est également intéressante lors d'une extension ou d'une reformulation de la requête, afin d'améliorer la correspondance entre la réponse et les besoins de l'usager.

3.1. La fonction d'indexation

Nous ne nous intéresserons dans cette présentation qu'aux différentes phases d'une composante d'indexation automatisée, bien que de nombreux systèmes utilisent une indexation manuelle ou une indexation assistée. Une bonne comparaison des différentes méthodes existantes se trouve dans [KERK 84].

La composante d'indexation regroupe les traitements nécessaires à l'analyse des documents et au stockage des résultats. Son objet est de produire une représentation du contenu de chaque document, conformément à un formalisme sémantique prédéfini, et en se restreignant aux notions

caractéristiques du document. C'est cet aspect restrictif de l'indexation qui va permettre au système de ne fournir en réponse à une requête que les documents jugés les plus pertinents.

Nous pouvons décomposer cette composante d'indexation en trois phases principales successives pour le traitement d'un document :

- une phase d'analyse textuelle
- une phase de normalisation et de sélection
- une phase de stockage

Nous allons exposer brièvement les fonctionnalités de ces différentes phases, afin de situer le cadre de notre travail

3.1.1. L'analyse textuelle

Comprenant les traitements linguistiques des documents, cette phase consiste en une analyse morpho-syntaxique (éventuellement complétée : analyse sémantique, pragmatique), qui doit permettre une extraction de termes ou de structures, susceptibles d'indexer le texte d'un document (i.e. de décrire les concepts représentatifs contenus dans un document).

L'analyse morphologique réalise la segmentation d'une phrase en morphes (mots). Ces morphes sont reconnus par cette analyse lorsqu'ils peuvent être décomposés en racines et désinences. Une ou plusieurs catégories grammaticales leur sont attachées avec les attributs grammaticaux associés (genre, nombre, personne, temps, ...). Des dictionnaires de racines, des modèles et des règles de décomposition morphologiques sont utilisés.

L'analyse syntaxique résout les ambiguïtés grammaticales (homographies) et regroupe les mots en syntagmes (groupe de mots constituant une unité à l'intérieur de la phrase), à l'aide d'un ensemble de règles constituant une grammaire de reconnaissance de la langue analysée.

Certains systèmes procèdent ensuite à une analyse sémantique dont le but est de résoudre les ambiguïtés sémantiques élémentaires (polysémies) et les ambiguïtés de structure résiduelles (bon agencement des syntagmes), en se référant à des connaissances qui font partie du système. Ces connaissances sont généralement spécifiques à un domaine particulier relatif au corpus considéré; Définies a priori, elles sont intégrées au niveau de dictionnaires (sous forme d'attributs sémantiques) et de bases de connaissances ou thésaurus (sous forme de relations sémantiques entre concepts).

On trouve également, en complément de ces analyses, des composantes pragmatiques, difficilement formalisables (solutions ad hoc), qui sont spécifiques aux domaines traités, et destinées à résoudre certains problèmes restés sans solution satisfaisante avec les analyses précédentes (généralement des ambiguïtés de nature syntaxique ou sémantique).

C'est cette phase de traitements linguistiques de la composante d'indexation, et plus particulièrement la partie analyse morpho-syntaxique, qui constitue l'objet principal de notre travail. Nous ne perdons toutefois pas de vue les contraintes imposées par les phases de sélection et de normalisation des concepts, de stockage des résultats, et celles relatives à la composante d'interrogation, qui nous ont conduit à effectuer certains choix, qui seraient discutables en dehors du contexte particulier de cette application.

3.1.2. La normalisation et la sélection

Cette phase a pour but la normalisation et la sélection des termes ou structures qui seront retenus comme termes d'indexation, parmi ceux produits (reconnus et extraits) par la phase d'analyse précédente. La complexité de ces structures et le niveau de sens associé sont fortement dépendants de la puissance de l'analyse mise en oeuvre. Un terme d'indexation étant tout ou partie d'un syntagme, véhiculant l'essentiel du sens (de l'information conceptuelle, du thème exprimé) d'une portion de texte.

La normalisation des concepts est une opération indispensable, pour pouvoir représenter, et donc par la suite identifier de manière univoque un même concept exprimé sous des formes différentes.

Cette normalisation peut s'opérer tout d'abord sur des critères morphologiques : lemmatisation consistant à ne considérer qu'une seule forme pour une famille lexicale. La lemmatisation peut être réalisée directement lors de la phase d'analyse précédente. La normalisation peut ensuite s'opérer sur des critères syntaxiques : ordonnancement des constituants de syntagmes suivant un ordre déterminé. Des considérations sémantiques ou pragmatiques peuvent enfin être prises en compte en accédant au thésaurus, par l'utilisation de relations sémantiques d'équivalence de type synonymie.

Les critères de sélection varient suivant les systèmes. Les principaux sont les suivants :

- sémantiques (comparaison avec les concepts présents dans le thésaurus),

- syntaxiques (structures syntaxiques particulières : portions plus ou moins complètes de syntagmes nominaux, de syntagmes verbaux, etc...),

- statistiques (fréquences d'apparitions dans le texte, dans un ensemble de textes du domaine, pondérations, etc...),

- structurels (nature de l'entité textuelle contenant le terme, utilisation de la structure logique du document, etc...).

Une stratégie de sélection peut combiner plusieurs d'entre eux.

Les termes d'indexation, une fois normalisés et sélectionnés, sont consignés dans une base d'indexation. Cette base d'indexation est généralement intégrée au thésaurus.

3.1.3. Le stockage

La phase de stockage des résultats a pour but la consignation des informations résultant des phases précédentes de manière à en optimiser l'exploitation.

Une relation d'indexation est établie entre les termes d'indexation sélectionnés et les documents référencés. C'est cette relation qui permettra ultérieurement, à l'issue du traitement de la requête d'interrogation, de fournir en réponse les documents jugés pertinents par le système.

Parmi les différentes méthodes utilisées pour la représenter, nous pouvons citer la technique du fichier inverse, qui est une liste de termes d'indexation, dans laquelle chaque terme d'indexation est associé à l'ensemble des références aux documents qu'il indexe. Pratiquement un fichier inverse peut être réalisé à l'aide d'une matrice M, où les lignes représentent des documents (la ligne i désigne le $i^{ème}$ document de la base, d_i), et où les colonnes représentent les termes d'indexation (la colonne j désigne le $j^{ème}$ terme d'indexation du domaine, t_j). Si le terme d'indexation t_j indexe le document d_i, alors $M(i,j) = 1$, sinon $M(i,j) = 0$.

Cette technique des fichiers inverses présente l'avantage de pouvoir supporter un grand nombre de documents (quelques millions), et est donc la plus utilisée.

En ce qui concerne la consignation et la structuration de la base d'indexation, ou du thésaurus, on peut utiliser les formalismes classiques de représentation des connaissances dans le domaine de l'intelligence artificielle : "frames", "scripts", réseaux sémantiques, règles de production, logique du 1er ordre, etc...

3.2. La fonction d'interrogation

La composante interrogation regroupe les traitements nécessaires à l'analyse et à l'interprétation des requêtes d'interrogation.

Il est à noter que le processus d'analyse nécessaire pour le traitement des requêtes n'est pas tout à fait le même que pour le traitement des documents, même si l'objectif de reconnaissance des termes exprimés est a priori similaire :

- dans le cas de l'indexation, on analyse un texte relativement long en langue naturelle, avec toute la complexité que cela implique. D'où la nécessité d'un processus performant, qui "avale" du texte.

- dans le cas de l'interrogation, l'analyse doit être plus fine, faisant appel si besoin est, à des mécanismes d'inférence, des consultations de thésaurus, pour obtenir une interprétation de la requête la plus correcte possible. Cela est facilité par le fait que les textes analysés sont courts et exprimés dans un formalisme souvent très restrictif par rapport à la langue naturelle.

Une fois l'analyse réalisée, l'interprétation d'une requête d'interrogation va s'effectuer classiquement par la production d'une équation de recherche, qui est un ensemble de termes d'indexation connectés par des opérateurs logiques. Les opérateurs les plus utilisés sont les opérateurs booléens : ET, OU, SAUF.

Par l'intermédiaire de la relation d'indexation, on détermine pour chaque terme d'indexation, son ensemble de références correspondant. Les opérateurs logiques sont ensuite appliqués sur ces ensembles de références, pour produire la réponse associée à la requête.

Il est intéressant ensuite d'essayer d'évaluer la pertinence de cette réponse, et éventuellement de reformuler (automatiquement ou de manière assistée) la requête.

L'architecture classique d'un système expert utilisant une base de connaissances et un moteur d'inférences est un point de départ intéressant pour traiter ces problèmes. Nous en verrons un exemple lors de la présentation du prototype IOTA développé au sein de notre équipe [CHIA 86 et 87].

3.3. Le thésaurus

La partie représentation et structuration de la connaissance relative à un domaine (construction d'un thésaurus) est un des aspects essentiels de la conception des systèmes de recherche d'informations. La dépendance avec les différentes phases de la composante indexation et avec la composante interrogation est évidente. Le rôle de la base de connaissances dans un SRI, apparait d'ailleurs en filigrane tout au long de la description sommaire que nous venons d'effectuer.

Un thésaurus est un ensemble de termes reliés par des relations sémantiques. D'une application à une autre, les termes présents dans un thésaurus varient énormément en fonction du domaine traité et du niveau et de la précision choisis pour la représentation des concepts (mots simples, groupes de mots, etc...). Par contre, les relations sémantiques utiles sont relativement bien connues, et leur utilisation est très semblable dans les différents systèmes [RIJS 79], [KERK 81], [BRUA 83], [SALT 83]. Ces relations peuvent être de caractère universel (valides pour toute application) ou de caractère contextuel (spécifiques au domaine traité par le SRI) :

- La relation de généricité :

  le terme *fruit* est un terme générique de *pomme*.

- La relation de spécificité (fonction inverse de la relation de généricité) :

  le terme *pomme* est un terme spécifique de *fruit*.

- La relation de synonymie pure,
  très rare dans la langue, elle apparait par exemple entre les sigles et leur définition :

  AFCET est synonyme de *Association Française pour la Cybernétique Economique et Technique*.

- La relation de synonymie partielle (ou contextuelle),
cette synonymie est valide dans un contexte sémantique particulier :

le terme *tableau* est synonyme de *matrice* en mathématiques,

ce qui n'est pas vrai dans d'autres domaines.

- La relation de voisinage sémantique,
cette relation indique une proximité sémantique entre deux termes, pour lesquels une relation de synonymie ne peut être utilisée :

le terme *habitat* est voisin-sémantique de *logement.*

- etc...

L'utilisation de ces relations sémantiques lors de la phase d'indexation, permet une normalisation des termes d'indexation :

- utilisation des relations de synonymie pour représenter par un seul terme d'indexation plusieurs expressions d'un même concept.

Lors de la phase d'interrogation, cet aspect normatif est toujours présent, mais ces relations servent également en cas de reformulation de la requête :

- utilisation de la relation de généricité, ou de voisinage sémantique pour étendre la portée de la requête,
- utilisation de la relation de spécificité, pour réduire cette portée.

Les thésaurus sont généralement constitués manuellement par des spécialistes du domaine d'application. Une comparaison des différentes méthodes utilisées pourra être trouvée dans [KERK 84]. Les coûts élevés et la non exhaustivité en restent les problèmes principaux.

Des outils de construction automatique (ou assistée), permettant l'établissement de relations entre les termes, sont développés, mais leur interprétation sémantique reste difficile à expliciter en termes conventionnels (synonymie, généricité, etc...). Nous verrons un exemple de construction et de structuration automatique d'une base de connaissances ayant donné des résultats prometteurs lors de la présentation des travaux développés au sein de notre équipe.

4. Les systèmes classiques

Nous ne présentons dans cette section que la partie indexation des systèmes classiques, utilisant une analyse linguistique. Nous renvoyons le lecteur à une étude spécifique aux phases d'interrogation développée dans [DEFU 86].

Les premiers systèmes de recherche d'informations réalisés ne comportaient pas de phase d'indexation automatisée. Les différents traitements relatifs à l'indexation (analyse, normalisation, sélection) étaient réalisés manuellement par des spécialistes du domaine ou des documentalistes. Outre les facteurs temps et coût, l'inconvénient principal de ces systèmes était d'ordre qualitatif : problème de cohérence lié à la subjectivité humaine et aux différents niveaux d'expertise (connaissance du domaine traité) des "indexeurs".

Des outils automatiques d'aide à l'indexation furent proposés. Parmi les différentes variantes développées, les principales consistaient en une comparaison des mots du texte avec :

- des antidictionnaires (dictionnaires de mots vides), permettant d'éliminer les mots non porteurs sémantiquement (n'exprimant pas de connaissance utile).
- des dictionnaires de termes d'indexation pré-définis pour le domaine d'application du corpus, permettant de les reconnaître dans les textes.

Ces différents dictionnaires sont constitués a priori pour un domaine particulier, soit manuellement, soit par intersection de textes représentatifs des domaines concernés.

Les méthodes d'indexation automatique développées se basent essentiellement sur deux modèles théoriques :

- le modèle statistique,
- le modèle linguistique.

Les méthodes statistiques ont été les premières a être utilisées dans le domaine de la recherche d'informations. Classiquement, leur objectif est d'évaluer l'importance d'un concept (généralement réduit à un mot), par rapport au contenu d'un texte, en fonction de sa fréquence d'apparition dans ce texte, et de sa fréquence d'apparition dans le corpus (ou plus généralement dans la langue).

Les systèmes basés sur ces méthodes statistiques produisent deux types d'indexation :

- une indexation binaire ou sélective :
  un terme indexe ou n'indexe pas un texte, comme dans les systèmes PASSAT ou MISTRAL [PASS 72] et [MIST 78],
- une indexation pondérée :
  le poids associé à un terme reflète son importance par rapport au texte, comme dans le système SMART [SALT 71].

Les méthodes purement statistiques pèchent par une carence au niveau de la résolution des problèmes linguistiques que sont :

- la normalisation du vocabulaire,
- la reconnaissance de concepts constitués par un ensemble de mots de la langue (problème des mots composés, par exemple : *pomme de terre*),
- la prise en compte de relations sémantiques entre les mots, de type synonymie ou généricité, et la résolution des références (pronominales, anaphoriques), dont l'ignorance fausse les calculs de fréquence, fondement de la méthode.

Elles sont donc complétées par des outils linguistiques (manuels ou automatiques), qui sont de deux types :

- des outils morpho-syntaxiques,
  dont le rôle classique est la normalisation du vocabulaire et la reconnaissance des concepts du texte,
- des outils sémantiques,
  dont le rôle est l'exploitation des relations sémantiques des mots de la langue pour affiner les concepts et lever certaines ambiguïtés résiduelles de l'analyse morpho-syntaxique.

Cette nécessité de traitements linguistiques qualitativement performants, pour obtenir une indexation satisfaisante, est maintenant reconnue unanimement dans le domaine [KERK 86], [SMEA 86 et 88]. Un des premiers systèmes de recherche d'informations utilisant ce genre de traitements linguistiques a été SYNTOL [BELL 70].

Ces extensions ont progressivement amené à augmenter la part des traitements linguistiques, indispensables à une meilleure définition du contenu des documents. Cette évolution a conduit à l'avènement de modèles dits linguistiques, dont la caractéristique essentielle est une fonction de correspondance requête-document fondée sur des critères de proximité linguistique assez poussés. Dans de tels systèmes, les calculs statistiques sont néanmoins présents pour traduire (une fois la correspondance linguistique

établie), la notion de pertinence relative des documents retrouvés. Un bon exemple de cette approche est le système SPIRIT [FLUH 81 et 85].

Une discussion plus développée de ces différentes approches peut se trouver de manière synthétique dans [KERK 84]. Une présentation détaillée peut se trouver dans [BOOK 75], [COOP 78], [RIJS 79], [DESC 82], [SALT 83, 86 et 87].

5. Le prototype IOTA

L'équipe Systèmes Intelligents de Recherche d'Informations, du Laboratoire de Génie Informatique de Grenoble, animée par Yves Chiaramella, est un groupe qui s'intéresse au domaine de la recherche d'informations. Les réalisations ont permis le développement d'un système prototype de recherche d'informations "intelligent" IOTA, composé à l'heure actuelle de trois modules :

- un module d'indexation automatique

- un module de constitution et de structuration automatique d'une base de connaissances

- un module "intelligent" d'interrogation

Outre la composante linguistique du module d'indexation automatique, qui fait l'objet de ce travail, d'autres études sont menées actuellement, afin de compléter ou d'améliorer les traitements mis en oeuvre dans les modules "opérationnels" [JIME 85 et 89] et [NIE 85 et 90].

5.1. Le module d'indexation automatique

Ce module développé par D. Kerkouba [KERK 84, 85 et 86], présente les caractéristiques suivantes :

- La définition d'une stratégie globale d'indexation, fondée sur l'exploitation de la structure logique complète du document (découpage en entités textuelles : chapitres, sous-chapitres, etc...), et sur l'exploitation d'éléments textuels informatifs (titres-procédés tels que les termes introductifs introduction et conclusion, titres informatifs, typologies particulières de certains termes du texte tels que les termes mis entre apostrophes, etc...).

Dans cette approche, le corpus est vu comme l'ensemble E des sous-arbres de la structure logique des documents, et la relation d'indexation R comme :

$$E \times C \times M \supseteq R$$

où C est l'ensemble des termes d'indexation, et M une mesure définie sur l'intervalle [0,1].

- L'ensemble des termes d'indexation est défini comme un sous-ensemble des syntagmes nominaux (nous reviendrons sur ce point au troisième chapitre), ce qui permet une représentation des concepts possédant un niveau de structuration proche de celui dans lequel sont exprimés ces concepts dans la langue naturelle. Ceci assure une meilleure adéquation entre la forme indexée du texte et son contenu, qu'en utilisant une représentation classique par mots simples (mots-clés). Les termes d'indexation sont donc des syntagmes nominaux extraits du texte, et normalisés par référence à une base de connaissances.

Le processus de construction des termes d'indexation est fondé sur des critères syntaxiques, et notamment sur un mécanisme de "cassure syntaxique", permettant de décomposer un concept initial (tel qu'il est extrait du texte), ne correspondant pas à un élément de la base de connaissances, en éléments plus courts qui seront à leur tour proposés pour la normalisation.

Cette approche a permis de définir une stratégie d'indexation dynamique basée sur la notion d'unité d'indexation minimale, définie en fonction du type de document. Les unités d'indexation minimales correspondent à des sous-arbres feuilles de l'arbre décrivant la structure logique d'un document.

Un mécanisme de "remontée automatique des termes d'indexation" dans la hiérarchie du document, permet, lors de la phase d'indexation, de déterminer le niveau de pertinence maximale pour un terme d'indexation dans la hiérarchie du document. Cette stratégie permet une meilleure

précision des résultats, en fournissant comme réponse à une requête la ou les entités textuelles les plus pertinentes, correspondant à un ou des sous-arbres de la structure logique du document.

- La relation d'indexation est pondérée par une mesure M basée sur la fréquence du terme d'indexation dans les diverses entités d'indexation du corpus. Cette mesure est définie à partir de l'évaluation statistique de deux notions complémentaires :

  * la représentativité d'un concept dans une entité textuelle, qui permet d'évaluer le degré de participation de ce concept dans l'information véhiculée par l'entité textuelle, compte tenu de l'ensemble des concepts évoqués dans le document;

  * la représentativité d'une entité textuelle par rapport à un concept, qui permet d'évaluer le degré de description du concept dans l'entité textuelle, compte tenu de sa description dans la totalité du corpus.

  Cette mesure dont le résultat est dans l'intervalle [0,1], évalue la représentativité mutuelle concept - entité textuelle.

Des exemples détaillés d'extraction de termes d'indexation à partir d'une syntaxe simplifiée des groupes nominaux (les traitements linguistiques faisant l'objet de notre travail n'étant pas encore disponibles au moment du développement de cette étude) et de stratégie de remontée de ces termes peuvent être trouvés dans [KERK 84].

5.2. Le module de constitution et de structuration automatique d'une base de connaissances

Ce module développé par M.F. Bruandet [BRUA 83, 85, et 87], permet de définir une base de connaissances à partir de l'analyse directe d'un ensemble de textes représentatifs d'un domaine. L'objectif essentiel est d'obtenir un thésaurus reflétant aussi fidèlement que possible les notions effectivement contenues dans le corpus traité, ce qui présente le double intérêt de limiter la taille de la base et son coût de constitution.

Les concepts de base sont représentés par des groupements de mots qui, par construction, constituent un sur-ensemble des groupes nominaux. Ces classes sont obtenues par extraction des cliques dans un graphe représentant les

liaisons de co-occurrence dans une même phrase des termes significatifs (i.e. susceptibles de constituer des groupes nominaux) dans le corpus de départ. Le graphe est lui-même défini par une matrice terme-terme dont les éléments contiennent une mesure définie sur l'intervalle [0,1] exprimant la force de la liaison contextuelle de deux termes. Cette mesure croît avec la fréquence de co-occurrence des deux termes, et est inversement proportionnelle à leur distance dans les phrases. Le choix des sous-graphes complets maximaux comme représentant de classes de concepts est une convention; ces classes regroupent par définition, des mots fortement liés contextuellement.

De par leur construction, les cliques peuvent être considérées comme l'expression de concepts de plus haut niveau trouvés dans le texte. Si l'on compare l'ensemble des cliques contenant un substantif particulier, on peut en déduire son contexte sémantique dans le corpus.

Exemple :

Voici un exemple de cliques extrait de [BRU 85], contenant le mot "support"; cet exemple est tiré d'un corpus d'expérimentation constitué par les NEF (manuel de normalisation pour les autocommutateurs, utilisé par les ingénieurs du CNET) :

(support microfiche), (support type de), (support microfilm être de), (support date de), (support information sur) , (support papier être de),

que l'on peut représenter graphiquement par la figure suivante, définissant l'environnement de "support"

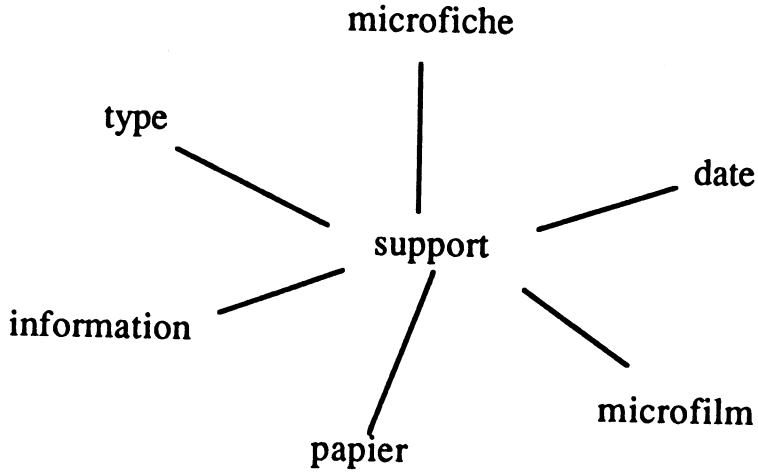


Figure 2 : *Représentation graphique d'une clique*

De plus pour un concept fortement représenté dans le texte, chaque clique le contenant définit une interprétation du terme considéré en donnant en quelque sorte ses différentes connotations. On peut se fonder sur ces propriétés pour mettre en évidence des relations sémantiques entre concepts (synonymie, etc...), dans le cadre d'un environnement assisté de construction de thésaurus : on peut considérer que deux termes sont sémantiquement liés s'ils ont beaucoup d'éléments communs dans leur environnement sémantique respectif.

C'est cet ensemble de cliques (sans les relations sémantiques) qui a servi de base de connaissances normalisatrice pour le module d'indexation automatique décrit précédemment [KERK 86].

5.3. Un module "intelligent" d'interrogation

L'objectif de ce module, outre la partie interrogation classique (traitement de la requête), consiste à simuler le comportement d'un expert humain en recherche d'informations. L'idée directrice est d'utiliser les apports de l'intelligence artificielle pour aborder les tâches nécessitant des traitements à un niveau plus sémantique (outils déductifs, apprentissage, etc...). Cette approche relativement récente a donné lieu a de nombreux travaux dans le domaine de la recherche d'informations : [SPAR 78], [SMIT 80] [BART 85] et [CROF 85 et 86].

L'étude menée par B. Defude a donné lieu à la réalisation d'un prototype de système expert en recherche d'informations [DEFU 86], [CHIA 86 et 87], l'approche choisie étant celle des systèmes experts procéduraux. Le moteur d'inférence est de type 0^+ (sans variables) avec un fonctionnement en chaînage avant et un mécanisme de retour arrière.
Les données des modules experts sont réparties en trois bases :

- une base de connaissances contenant les règles correspondant aux tâches expertes et au processus de coopération (système expert - SRI),
- une base de données à court terme,
- une base de procédures.

Le processus de traitement d'une requête peut être vu comme l'exécution séquentielle de différents modules, pouvant être qualifiés d'expert ou non, selon un plan d'action géré par le système expert (coopération). Schématiquement, l'enchaînement est le suivant :

-1) analyse de la requête.

La requête se présente (pour l'instant) comme une expression booléenne de groupes nominaux. Cette phase comporte éventuellement un traitement des mots inconnus.

-2) processus de correspondance concepts - termes d'indexation.

Ce traitement est fondé sur un processus de pattern-matching syntaxique analogue à celui utilisé dans l'indexation, et aboutit à l'équation finale de recherche.

-3) évaluation du niveau de dégradation de la requête et de la typologie de l'utilisateur.

Les données sont nécessaires aux modules experts, pour diriger notamment les stratégies d'évaluation et de reformulation des réponses.

-4) interprétation de l'équation finale de recherche.

Cette équation finale est en fait une expression booléenne de termes d'indexation. Le traitement consiste à rechercher les références correspondantes par consultation de la relation d'indexation.

-5) évaluation du résultat.

Elle consiste à évaluer globalement la qualité des références retrouvées, au vu notamment de la typologie de l'utilisateur (tâche typiquement experte) :

satisfaisant --> (7) sinon --> (6)

-6) Reformulation.

La reformulation vise à rechercher une réponse plus satisfaisante compte tenu de la qualité de la réponse précédente, et de la typologie de l'usager. Elle consiste à engendrer une nouvelle requête par modification des concepts et/ou des opérateurs logiques, selon le but assigné

--> (3)

7) Apprentissage des inférences réalisées.

Les points les plus intéressants et les plus importants de ce système sont, d'une part, l'évaluation de la typologie de l'utilisateur, qui n'est faite qu'une seule fois à partir de la première équation finale de recherche, et qui mesure le niveau de l'utilisateur dans le domaine (un spécialiste n'a pas les mêmes besoins en information qu'un débutant, ce qui entraîne des traitements différents de la requête), et d'autre part, l'évaluation du résultat qui conditionne le cas échéant, la reformulation automatique de la requête.

Ces tâches sont typiquement du ressort d'un expert en recherche d'informations. Les connaissances expertes sont stockées sous forme de règles de production :

Si < condition > Alors < action >

Voici un exemple d'une telle règle pouvant servir pour la partie évaluation du résultat en fonction de la typologie de l'utilisateur :

SI tâche-courante = "évaluation des réponses"
ET typologie-utilisateur = "spécialiste"
ET nombre-réponses > 25
ALORS
évaluation-réponses := "insatisfaisant"
tâche-courante := "reformulation"

La partie gauche d'une telle règle est une condition portant sur des variables d'états du système, qui se trouvent dans une base de donnée à court terme qui contient aussi la pile des buts à accomplir, et qui joue le rôle de "tableau noir" pour les différents modules activés.

La partie droite est une séquence d'actions permettant en plus des mises à jour de la base à court terme (état des variables, données intermédiaires, etc...), l'appel de procédures externes (stockées dans la base des procédures).

6. Conclusion

Il est bon de préciser que nous employons le terme de "Système de Recherche d'Informations", SRI, de préférence à "Système de recherche documentaire", terme qui peut paraître plus restrictif ou plus archaïque, mais dont la traduction anglo-saxonne est la même : "information retrieval systems". Ce changement de dénomination correspond à un élargissement considérable du domaine, à la fois au niveau des fonctionnalités (on ne s'intéresse plus exclusivement aux bases de données bibliographiques), et au niveau des modèles sous-jacents (la tendance actuelle étant la définition de modèles beaucoup plus sophistiqués, dont la mise en oeuvre est fondée sur l'intégration de techniques issues de divers domaines : les bases de données, l'intelligence artificielle, les traitements automatiques de la langue naturelle, etc ...).

Nous nous sommes restreints volontairement dans cette présentation, aux systèmes de recherche d'informations permettant de gérer et d'exploiter des bases de documents uniquement textuels, bien que l'évolution actuelle vise à l'intégration de documents hétérogènes, qualifiés de "multi-médias", et pouvant comprendre, outre du texte, des images et des sons [CHRI 87]. Un ensemble de documents regroupés dans une base constitue un corpus relatif à un domaine ou à un ensemble de domaines particuliers.

Il est à noter qu'actuellement, pour le développement de SRI appliqués à des bases de documents textuels, de nombreux problèmes ne sont que partiellement résolus ou même abordés, notamment au niveau des capacités d'apprentissage (capacités d'évolution automatique intelligente du système), et de la définition du niveau de structuration des connaissances extraites automatiquement. C'est tout le sens de notre étude que d'essayer d'y apporter un certain nombre d'améliorations ne mettant pas en oeuvre de mécanismes trop complexes (qui les rendraient impraticables sur des applications de taille réelle), que nous développerons dans les chapitres suivants.

Notre travail s'inscrit dans le cadre du développement d'un système intelligent de recherche d'informations, et concerne la composante linguistique du processus d'indexation automatique qui comme on l'a déjà vu, ne peut donner de bons résultats sans une telle composante [SMEA 88].

L'expérimentation du processus d'indexation automatique développé par D. Kerkouba a permis de vérifier cette nécessité, en mettant en évidence les interprétations erronées obtenues au niveau de la normalisation et de la reconnaissance la plus élémentaire du vocabulaire (utilisation d'un lemmatiseur). De nombreux biais sont introduits à cette occasion qui faussent en partie les résultats de l'indexation [KERK 86].

L'objectif de ces traitements est de fournir au processus de sélection, des concepts bien reconnus, extraits des textes analysés, et susceptibles de constituer des termes d'indexation.

Nous avons vu qu'une condition d'une bonne adéquation entre le texte analysé et son contenu était une représentation des concepts comportant un niveau de structuration proche de celui dans lequel sont exprimés les concepts dans la langue naturelle, ce qui nous a conduit à nous intéresser plus particulièrement aux syntagmes nominaux. Nous justifierons ce choix lors de la présentation des Groupes Conceptuels dans le chapitre III. Nous verrons dans le chapitre suivant, lors de la présentation de certains systèmes d'analyse du français développés ou appliqués dans le cadre des systèmes de recherche d'informations, que la correspondance groupe nominal - terme d'indexation est largement admise, et influence les traitements mis en oeuvre. Un aspect

important lié à ce cadre d'application est la nécessité de définir des outils d'analyse performants, car destinés à traiter des corpus volumineux.

Enfin l'aspect évolutif de ces systèmes d'analyse (en particulier en ce qui concerne le vocabulaire) nous parait être essentiel, si l'on veut pouvoir appréhender des univers textuels ouverts. Nous verrons dans le chapitre suivant que l'utilisation de données linguistiques trop précises, trop expertes (au sens où l'intervention d'un expert est nécessaire) constitue une limitation à cet aspect évolutif. En conséquence nous essayerons de faciliter cette évolution en définissant un ensemble de données linguistiques suffisamment précises pour permettre une analyse intéressante, mais assez générales toutefois, pour permettre l'appréhension d'univers textuels ouverts sans nécessiter d'interventions expertes pour permettre l'évolution du système.

CHAPITRE II

PLAN DU CHAPITRE II

TRAITEMENT AUTOMATIQUE DE LA LANGUE NATURELLE

TRAITEMENT AUTOMATIQUE DE LA LANGUE NATURELLE

1. Introduction

Nous nous sommes restreints, après un bref rappel historique, à la présentation de cinq systèmes d'analyse du français écrit, ayant eu des applications dans le domaine des systèmes de recherche d'informations. Nous avons choisi ces systèmes, bien que certains soient relativement anciens, en raison de l'originalité des méthodes utilisées et des rapports que ces méthodes peuvent avoir avec nos travaux. Nous dégagerons de ce tour d'horizon quelques considérations qui sont à la base de notre démarche.

Il est utile, avant d'introduire les traitements linguistiques employés dans les systèmes de recherche d'informations, d'effectuer un bref rappel historique (n'ayant pas de prétention d'exhaustivité), destiné à mettre en évidence le cheminement ayant conduit aux diverses orientations de la recherche actuelle dans ce domaine.

2. Rappel

Les premières tentatives de traitement automatique des langues naturelles, au début des années 50, furent liées essentiellement au domaine de la traduction automatique. Les premiers systèmes fonctionnaient en effectuant une traduction mot à mot de la langue source vers la langue cible, sans analyse poussée du langage naturel. Traitant de domaines particuliers, une simple analyse morphologique suffisait à établir une correspondance de termes dans les lexiques des langues traitées. Les ambiguïtés grammaticales étaient résolues de manière très simple, généralement au niveau des lexiques; ces résolutions étant valable pour le domaine d'application concerné. Une correspondance entre règles de grammaire de la langue source et de la langue cible, permettait de présenter les résultats de manière compréhensible.

Après une période d'enthousiasme, le relatif échec de ces systèmes (limités à des domaines très particuliers, avec un vocabulaire restreint), mit en évidence la nécessité d'effectuer une analyse syntaxique complète de la langue naturelle afin de pouvoir obtenir une meilleure compréhension "automatique".

Dans le même temps se sont alors développés les travaux de théorie linguistique formelle basés sur la syntaxe structurale et les grammaires transformationnelles [TESN 59], [HARR 61 et 71], [CHOM 57 et 65], qui donneront naissance à un certain nombre de travaux privilégiant l'analyse syntaxique automatique des langues naturelles. Directement issu des travaux de Z.S. Harris, un analyseur syntaxique de l'anglais fut réalisé par N. Sager [SAGE 67]. M. Salkoff réalisa une grammaire en chaîne du français, destinée à être intégrée à cet analyseur [SALK 73]. Il s'agit en fait d'un français particulier, car la définition de la grammaire prend en compte le domaine d'application du corpus concerné, en l'occurrence les textes scientifiques (mathématiques, biologie, etc...), et certaines connaissances sémantiques sont nécessaires au système, notamment pour la détermination de l'affectation des sous-classes syntaxiques (qui pourrait être différente dans d'autres contextes).

Par ailleurs, l'émergence de l'intelligence artificielle essayant de simuler le raisonnement humain, a orienté des recherches vers des systèmes cherchant "à comprendre" le langage naturel (ou à simuler la compréhension) en traitant le "sens" véhiculé par les textes. Ainsi apparurent notamment, les premiers systèmes de Question-Réponse, fonctionnant simplement par reconnaissance de mots clés ou d'expressions clés dans les questions. La reconnaissance d'un mot clé provoquant l'activation d'une réponse stéréotypée. Le système ELIZA simulant le dialogue entre un psychiatre et son client en est un bel exemple [WEIZ 66]. Le relatif succès de tels systèmes encouragea les recherches dans le sens d'une plus réelle compréhension.

Des outils informatiques furent alors développés permettant (à des degrés divers) de combiner syntaxe et sémantique. Parmi les principaux formalismes nous pouvons citer les A.T.N. (Augmented Transition Network) de Woods [WOOD 70 et 80], les système-Q du projet TAUM [COLM 71], l'approche procédurale de Winograd avec son système SHRDLU [WINO 72], la dépendance conceptuelle de R.C. Schank [SCHA 72 et 75].

Plus récemment, l'avènement d'outils généraux tels que le langage de programmation PROLOG (PROgrammation en LOGique), basé sur la logique du 1er ordre restreinte aux clauses de Horn [COLM 83] est de nature à bouleverser le domaine. Ce langage est particulièrement bien adapté aux traitements des langues naturelles, dans la mesure où l'écriture d'un analyseur peut se faire "directement" à partir de l'écriture de la grammaire du langage à analyser, sous forme de grammaires logiques (telles que les grammaires de métamorphose, par exemple [COLM 75]); la transcription des règles de grammaires en clauses PROLOG étant immédiate. La programmation est déclarative, un programme est un ensemble de clauses qui représentent soit des assertions soit des règles. En utilisant le principe de résolution de J.A. Robinson [ROBI 65], les règles permettent de déduire par inférence des résultats à partir des assertions. L'aspect interprétatif et "combinatoire" du mécanisme de résolution (recherche de toutes les règles applicables, avec mémorisation de tous les choix possibles pour les retours-arrières ultérieurs), font de ce langage un outil intéressant pour les tests et études de faisabilité, mais le rendent actuellement inefficace pour des traitements de grande envergure, car beaucoup trop coûteux en place et en temps. Les développements actuels du langage (compilation, meilleure gestion de la combinatoire) permettront sans doute de réviser, à terme, ce jugement.

D'autres chercheurs se préoccupèrent de traiter directement le sens à partir du langage naturel. C'est ainsi qu'apparut le concept de "réseau sémantique" [QUI L 68], [SIMO 72] destiné à représenter, sous un formalisme approprié aux traitements informatiques, une certaine connaissance. En schématisant, on peut considérer un réseau sémantique comme un graphe dont les sommets sont constitués par des mots représentant des concepts, reliés par des arcs orientés et étiquetés représentant des relations sémantiques.

L'historique du traitement automatique des langues naturelles et la présentation détaillée des différents formalismes ayant longuement été présentés dans la littérature, nous ne le referons pas plus avant ici, et nous renvoyons aux nombreux ouvrages de synthèse, entre autres [JAYE 79 et 86], [WINO 83], [SPAR 83], [MELL 85], [PITR 85], [COUL 86], et [VAUQ 75] pour un rappel historique.

3. Quelques systèmes classiques pour le français

Nous allons présenter maintenant quelques travaux relatifs au traitement automatique des langues naturelles, réalisés pour le français. Le point commun de ces systèmes, outre de traiter le français, est d'avoir été utilisés dans des Systèmes de Recherche d'Informations.

Nous ne développerons donc pas ici des systèmes très intéressants comme le système ARIANE-78, développé au GETA de Grenoble (Groupe d'Etudes pour la Traduction Automatique), par une équipe animée par B.Vauquois, et actuellement par Ch. Boitet [BOIT 82], ou comme le système METEO, développé par le groupe TAUM (Traduction Automatique de l'Université de Montréal) [CHAN 76]. Ces systèmes relèvent du domaine de la traduction automatique, domaine où la problématique est très différente de celle de la recherche d'informations. Le développement de ces outils linguistiques répond en effet à des contraintes bien particulières; si de nombreux résultats d'études plus générales, ou relevant d'autres domaines d'application, sont fondamentaux et peuvent être réutilisés dans notre contexte, la problématique de la recherche d'informations (et plus particulièrement, pour ce qui nous concerne, de l'indexation automatique) implique le développement de méthodes spécifiques.

D'autres domaines tels que celui de la reconnaissance de la parole, avec en particulier les travaux réalisés au CRIN à Nancy par l'équipe "Reconnaissance des Formes et Intelligence Artificielle" animée par J.P. Haton et J.M Pierrel, utilisent des traitements sophistiqués de la langue naturelle. De même que le domaine de la correction automatique de textes, très porteur, car fondamental avec les procédés de lecture optique et les systèmes de traitement de textes, qui sont en plein essor actuellement.

On peut trouver dans la littérature de nombreuses présentations et analyses synthétiques de ces systèmes, nous citerons [ZAJA 86] pour la traduction automatique, [PERE 86] pour la correction automatique de textes, [PIER 82] , [DIVA 84] et [HATO 85] pour la reconnaissance automatique de la parole.

De la même manière, nous ne présenterons pas ici les travaux portant sur l'étude de la langue naturelle réalisées notamment au LADL à Paris (Laboratoire d'Automatique Documentaire et de Linguistique), par une équipe animée par M. Gross, réalisant des études systématiques portant sur un ensemble important de propriétés syntaxiques par rapport à un lexique [GROS 75 et 86]; et au centre de recherche pour un Trésor de la Langue

française de Nancy, (TLF), portant sur des études statistiques de vocabulaires réalisées à partir d'une collection importante de textes littéraires choisis des XIXème et XXème siècles, et contenant 70 millions de formes [TLF 71]. On n'insistera jamais assez sur l'importance de ces travaux, d'une part pour une meilleure connaissance du français, et d'autre part pour toutes les équipes qui s'essaient à traiter automatiquement la langue naturelle, quel que soit le domaine particulier d'application.

Nous allons commencer cette présentation par l'analyseur en chaîne de M. Salkoff, qui bien que relativement ancien, est le plus représentatif pour le français de l'ensemble des analyseurs inspirés directement des travaux de Z. Harris, [SALK 73 et 79]. L'intérêt de ce travail réside essentiellement dans la description très détaillée des syntagmes nominaux, auxquels nous nous intéressons tout particulièrement. La limitation principale de cette démarche nous semblant être l'introduction dans le lexique de renseignements linguistiques d'ordre syntaxico-sémantique, tels que les propriétés rectionnelles (qui constituent des canevas syntaxiques pour les formes régissant des compléments), ou les sous-catégorisations sémantiques (rappelant les "traits sémantiques" de J.J. Katz et J.A. Fodor [KATZ 64]). Ces données linguistiques, forcément déterminées a priori par des spécialistes, sont utilisées comme des contraintes entre les éléments d'une même phrase. Ce type d'approche donne de bons résultats sur un domaine bien cerné, où le vocabulaire utilisé est connu et peut être défini a priori, mais demande un gros investissement dès que l'on veut étendre le langage.

Le programme de segmentation automatique du français réalisé par B. Maegaard et E. Spang-Hanssen [MAEG 78], que nous présentons ensuite, se particularise par l'utilisation de moyens relativement simples et efficaces, pour réaliser une analyse de surface de la langue naturelle. Cette approche s'apparente assez à la nôtre, même si les objectifs poursuivis ne sont pas les mêmes : segmentation automatique des phrases en propositions (l'effort essentiel porte dans ce cas sur la reconnaissance des groupes verbaux). Sa caractéristique essentielle est le repérage d'un ensemble de marqueurs ou de formes bien précises du texte, exploités ensuite par un automate de segmentation. La résolution d'une ambiguïté grammaticale ne sera tentée par un examen du contexte proche, que dans la mesure où elle pourra contribuer à ce repérage. Ce point nous parait fondamental, et constitue une des hypothèses de base de notre approche. La stratégie consistant a réaliser une analyse de surface des textes produisant un résultat traitable ensuite, par un automate réalisant les objectifs finaux, constitue également un point commun avec notre démarche. Cette séparation analyse linguistique de surface - application permet d'envisager une certaine généralité de ces méthodes. La récupération de cette segmentation est d'ailleurs envisagée dans le cadre du projet SYDO que nous présentons plus loin, les modèles linguistiques utilisés relevant dans les deux systèmes de la même philosophie.

Le système du projet SYDO a été spécialement conçu pour le domaine de la recherche d'informations [ROUA 83], [LALL 86] et [KALL 87], et si certains aspects de ce système rejoignent nos préoccupations (une certaine classification des ambiguïtés grammaticales, le repérage d'indicateurs de structures, un intérêt particulier pour les groupes nominaux), l'approche est radicalement différente de la nôtre, tant sur les aspects outils (utilisation d'un filtre statistique, mise en oeuvre d'une analyse syntaxique poussée), que sur les aspects architecture (les traitements morphologiques et syntaxiques sont très compartimentés et nécessitent d'ailleurs des prétraitements importants). Là encore, une sous-catégorisation basée sur des critères syntaxico-sémantiques, nous parait un obstacle important, tout comme dans le cas de l'analyseur en chaîne de M. Salkoff, à une généralisation des domaines d'application.

Nous présentons également le système PIAF [COUR 77], qui est un système interactif d'analyse syntaxique du français, ayant donné lieu à de nombreuses applications dans des domaines différents, et notamment dans le domaine de la recherche d'informations, avec les logiciels PIAFDOC [GRAN 80], PIAFPS [MERL 82]. Ce système d'analyse est très puissant, mais repose sur l'interactivité, qui permet de s'en remettre à l'utilisateur pour résoudre un certain nombre de problèmes, tels que l'apparition de mots nouveaux dans un texte et la résolution des ambiguïtés. Cela implique une certaine connaissance du système de la part de l'usager, et provoque de nombreuses interruptions du processus d'analyse, déjà relativement coûteux (production de toutes les structures syntaxiques possibles d'une phrase correspondant à des interprétations différentes). En ce sens, le logiciel PIAFPS dont l'objectif est de réaliser une présyntaxe permet la réduction de ce coût.

Enfin nous présentons le système basé sur des algorithmes à apprentissage développé par l'équipe de A. Andreewsky, [ANDR 73 et 77], qui ont été à la base de la réalisation du logiciel SPIRIT [FLUH 81 et 85]. Les travaux de cette équipe ont dès le départ été orientés vers la réalisation de systèmes documentaires et ont donc de ce fait proposé des algorithmes spécialement conçus pour appréhender des univers textuels ouverts. Ce point est essentiel et a conduit cette équipe à s'intéresser aux algorithmes à apprentissage, afin d'acquérir automatiquement le maximum d'informations linguistiques [FLUH 77] et [DEBI 82]. Ces travaux ont permis de vérifier que l'utilisation du fort caractère positionnel de certaines langues naturelles, et en particulier du français, pouvait donner des résultats intéressants, avec un coût très raisonnable. Notre analyse repose en grande partie sur cet aspect positionnel du français, et un des outils mis en oeuvre (le filtrage positionnel et grammatical à l'aide d'une matrice de précédence) est directement inspiré de cette approche. Nous nous sommes toutefois démarqué de l'utilisation de méthodes statistiques, qui, si elles permettent de fournir dans tous les cas, au moins une interprétation (la plus probable), ne garantissent pas l'exactitude des résultats obtenus. Nous verrons que dans notre analyse, au contraire, nous nous

sommes attachés à n'utiliser et à n'acquérir que des données que l'on peut qualifier de certaines (quitte dans certains cas à être incomplètes).

3.1. L'analyseur en chaîne de M. SALKOFF

En collaboration avec l'équipe de N. Sager à New-York, qui s'est inspirée directement des travaux de Z. Harris sur la méthode des coupures successives [HARR 61] pour réaliser un analyseur syntaxique de l'anglais [SAGE 67], M. Salkoff réalisa une grammaire en chaîne du français [SALK 73 et 79] destinée à être intégrée à cet analyseur.

Une chaîne est une structure syntaxique constituée par une succession de catégories grammaticales entretenant des rapports grammaticaux entre elles.

L'idée directrice de l'analyse en chaîne est de supposer que chaque phrase peut être décomposée en une succession de chaînes imbriquées les unes dans les autres. La chaîne élémentaire qui n'est imbriquée dans aucune autre chaîne se nomme la chaîne centrale (elle constitue le squelette de la phrase), les autres sont des chaînes d'ajout. Dans la grammaire traditionnelle, les chaînes d'ajout constituent ce que l'on nomme les modificateurs.

Exemple :

dans la phrase

"Jean, qui était avec nous hier soir, regardait attentivement la télévision en mangeant "

on peut isoler la chaîne centrale :

" Jean regardait la télévision "

et les chaînes d'ajout :

" qui était avec nous hier soir ", " attentivement ", " en mangeant "

La chaîne centrale est le noyau minimum de la phrase auquel on ne peut plus rien enlever sans qu'elle cesse d'être une phrase.

Les chaînes d'ajout ont leurs structures propres et peuvent s'insérer dans la chaîne centrale, ou dans d'autres chaînes d'ajout, en des points déterminés : soit à gauche ou à droite d'un élément particulier de la chaîne, soit en n'importe quelle position dans la chaîne.

Dans l'exemple précédent, la chaîne d'ajout " *hier soir* " est greffée sur une autre chaîne d'ajout, mais aurait pu s'insérer en d'autres points de la phrase, que nous matérialisons par des "*" :

" * *Jean* *, *qui* * *était* * *avec nous* *, *regardait* * *attentivement* * *la télévision* * *en mangeant* *. "

Etablir une grammaire en chaîne consiste à définir de manière précise les différentes structures des chaînes centrales et des chaînes d'ajout classées en fonction de leurs points d'insertion (dA ajout à droite de l'adjectif, gN ajout à gauche du nom, etc...).

Voici quelques exemples simples de chaînes centrales et de chaînes d'ajout, tirés de [SALK 73] :

- La chaîne centrale d'assertion désignée par C1 :

$$C1 = \Sigma_i \, tV_{ij} \, \Omega_j$$

" *Pierre lit son livre* "

où Σ représente un sujet, tV un verbe fléchi, Ω un objet, et où les indices indiquent des contraintes entre sujet et verbe d'une part, et entre verbe et complément d'autre part.

- Les ajouts phrastiques à droite d'un adjectif :

de V Ω *que* C1

" *Las de lire son livre* ", " *certain que Pierre lit son livre* "

où V représente un verbe à l'infinitif. Il existe bien d'autres types possibles d'ajouts phrastiques à droite d'un adjectif que ceux cités dans cet exemple.

Un ensemble de règles de réécriture hors-contexte donne les structures des différentes chaînes par l'ordre des catégories y participant. Un ensemble de

restrictions sur ces chaînes traduisant des rapports grammaticaux entre les catégories de la langue et de contraintes portant sur des sous-classes grammaticales (Nh nom humain, etc..) permet d'éliminer les segmentations conduisant à des interprétations erronées :

Exemple :

Si l'on prend les deux verbes " *penser* " et " *déjeuner* " leur sujet est forcément un nom humain :

$$\Sigma = Nh$$

or, le verbe " *penser* " admet les deux types de compléments suivants :

$$\Omega = \emptyset \quad \text{dans} \quad \text{" } \textit{Pierre pense.} \text{ "}$$
$$\Omega = \textit{que C1} \quad \text{dans} \quad \text{" } \textit{Pierre pense que Paul lit son livre} \text{ "}$$

par contre si on peut avoir :

$$\text{" } \textit{Pierre déjeune} \text{ "} \quad \text{avec} \quad \Omega = \emptyset$$

on ne peut jamais avoir un complément en *que* C1 avec "*déjeuner* "

L'établissement de sous-classes sur des critères uniquement syntaxiques n'étant pas possible, des considérations d'ordre sémantique, liées au domaine d'application concerné, sont utilisées. Cet aspect enlève une certaine généralité à cette grammaire.

Pourtant un reproche fréquemment fait à M. Salkoff est le peu d'importance accordé aux traitements sémantiques dans son analyseur [RADY 83]. Nous pensons au contraire, que l'introduction de contraintes sémantiques dans une grammaire, lui enlève tout caractère de généralité et en limite l'utilisation à des applications dans un domaine particulier. M. Salkoff lui même, affirme que son analyseur contient déjà une certaine sémantique peut-être excessive, notamment au niveau des restrictions et des sous-classes [SALK 87].

L'analyseur qui en découle est fondamentalement descendant, et utilise systématiquement la technique du retour-arrière dès que l'analyse est dans une impasse.

L'effort de M. Salkoff dans l'écriture de cette grammaire a porté principalement sur la définition du groupe nominal, et ce travail constitue à notre connaissance l'étude la plus complète sur le sujet.

3.2. La segmentation automatique du français écrit

Ce programme de segmentation automatique du français écrit a été réalisé par une équipe de l'université de Copenhague, animée par B. Maegaard et E. Spang-Hanssen [MAEG 78].

Le but de cet analyseur est l'analyse de textes en propositions (principales, relatives, etc...). L'hypothèse de départ est qu'il est possible d'identifier le début et la fin des propositions en français, dans la grande majorité des cas, à l'aide d'un ensemble de marqueurs ou de formes bien précises :

- Les verbes finis (verbes à une forme personnelle).

- Les introducteurs de propositions subordonnées (conjonctions de subordination et pronoms relatifs).

- Les conjonctions de coordination.

- La ponctuation.

Parmi ces marqueurs, seuls les verbes appartiennent à une classe "ouverte", et leur identification n'est pas simple notamment à cause des nombreuses homographies possibles. En conséquence, l'intérêt s'est porté tout naturellement sur l'identification des syntagmes verbaux.

Le premier composant de cet analyseur consiste en une analyse morphologique (attribution de catégories aux mots du texte) à partir d'un modèle linguistique comprenant 10 catégories grammaticales (certaines ne concernant d'ailleurs qu'un très petit nombre de mots dont toutes les formes sont répertoriées dans des listes).

Seule l'identification des verbes parmi les classes "ouvertes" est recherchée de manière systématique, et le système dispose d'un lexique important (5000 racines verbales et 600 désinences). Les substantifs et la plupart des adjectifs sont considérés comme "non-marqués", ou neutres, et sont rangés dans une classe adéquate N, dans laquelle sera également rangé tout mot inconnu du système.

La résolution des homographies se fait par l'examen d'un contexte étroit des formes ambiguës à l'aide d'un ensemble de tests appropriés. Une batterie de six tests est proposée pour cette résolution, qui ne concerne que les homographies des formes verbales :

- S = test substantif (exemple : *porte*)
- J = test des participes passés en _s et des adjectifs (exemple : *mis, vide*)
- M = test des participes passés en _t (exemple : *dit*)
- N = test du participe passé *fait*
- K = test de la préposition *entre*
- L = test des locutions du type *à reculons*

Chaque forme verbale ambiguë contient au niveau du dictionnaire une indication sur le test qui devra lui être appliqué. Une étude minutieuse des formes verbales ambiguës présentes dans le dictionnaire est nécessaire pour la détermination des bons tests à associer (il est nécessaire de connaître l'ensemble des homographes pour une forme donnée, afin d'effectuer le bon choix).

Enfin la chaîne des catégories est réduite par regroupement : les pronoms personnels sont absorbés par les verbes associés, les catégories auxiliaires D (déterminant) et P (préposition) sont transformées en non-marquées N car elles ne sont plus utiles pour la suite du traitement.

Une fois cette analyse morphologique effectuée, le système réalise la segmentation en propositions à l'aide d'une grammaire en réseau, équivalente à une grammaire "context-sensitive", dont l'automate de reconnaissance associé est du type automate linéaire borné (automate à pile qui peut lire dans les deux sens sur la bande d'entrée, et qui peut lire et écrire partout sur la bande de stockage).

Exemple :

Voici un exemple d'analyse d'une phrase, extrait de [MAEG 78] :

" *Sur les marches de l'escalier qui tournait et qui était raide, elle lui demanda :* "

Cette phrase comporte une proposition principale et deux propositions subordonnées coordonnées. L'analyse morphologique a produit pour cette phrase la chaîne de catégories grammaticales suivante :

N N N N N N I V E I V N S W A

avec les conventions suivantes : N = non-marqué ou neutre, I = introducteur de proposition subordonnée, V = verbe fini sans pronom personnel sujet, E = conjonction de coordination,

S = virgule, W = verbe fini avec pronom personnel sujet, A = ponctuation forte.
On peut remarquer dans ce résultat que la catégorie W regroupe les trois formes " *elle lui demanda* " (absorption des pronoms par le verbe).

Le résultat fourni par l'automate de segmentation est :

N N N N N N/ W A
I V/E I V N S/

Le symbole "/" est une borne pour les propositions (marque la séparation entre deux propositions). Le décalage matérialise les différents niveaux d'enchassement des propositions.

L'intérêt de ce travail réside essentiellement, outre la segmentation en propositions proprement dite, en une suggestion de moyens relativement simples (batterie de tests) pour résoudre des ambiguïtés grammaticales, par exploration du contexte proche, sans avoir recours à une analyse syntaxique classique. Il s'agit en fait d'une analyse de surface du français.

Même si, du fait du petit nombre des catégories grammaticales et des objectifs visés, la résolution des homographies ne concerne uniquement que l'appartenance à la classe des formes verbales, la fréquence de ce type d'homographies dans la langue justifie pleinement une telle étude.

3.3. Le système P.I.A.F.

Le système P.I.A.F. (Programme Interactif d'Analyse du français) est un analyseur morpho-syntaxique du français, conçu pour produire des structures de dépendances. Une grammaire de dépendances est une grammaire générative faiblement équivalente à une grammaire "hors contexte" [COUR 77].

Exemple :
Une phrase telle que :

" *Le chauffeur conduit la voiture* "

est analysée en tant que "gouverneurs" et "dépendants" de la manière suivante :

(*chauffeur* (*le* *))

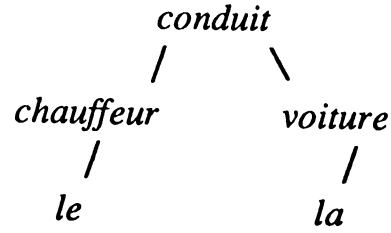
(*voiture* (*la* *))

indiquant que " *le* " est un dépendant du gouverneur "*chauffeur* " et que " *la* " est un dépendant du gouverneur "*voiture* ".
Le symbole "*" indiquant la place du gouverneur par rapport à ses dépendants.

Puis,

(*conduit* (*chauffeur* (*le* *) * *voiture* (*la* *)))

d'où la structure suivante :



Le système P.I.A.F. est composé de deux modèles principaux qui sont :
- un transducteur général d'états finis,
- un système d'analyse syntaxique.

Le transducteur général d'états finis a été conçu pour réaliser l'analyse morphologique d'une langue, qui consiste à découper une phrase en segments, et à effectuer une transduction qui produit les renseignements linguistiques associés aux segments (catégories et valeurs de variables grammaticales).

Un segment est déterminé par sa décomposition à l'aide d'un dictionnaire, en éléments nommés : préfixes, bases ou racines, suffixes, désinences. Le contrôle est effectué par l'application des règles d'une grammaire permettant de valider la concaténation de ces éléments, et donc de vérifier la bonne segmentation de la chaîne d'entrée. Le caractère "blanc" n'est pas considéré comme un séparateur de mots, de manière à pouvoir considérer un mot composé ou une tournure idiomatique comme un seul segment ("*au fur et à mesure*").

Les renseignements linguistiques comprennent toutes les interprétations de toutes les décompositions possibles d'un segment et seront utilisés par l'analyse syntaxique. La transduction produisant ces renseignements est réalisée par la même grammaire, qui est une grammaire à validations et à saturations dont l'équivalence à une grammaire d'états finis a été démontrée dans [COUR 77]. L'ensemble des validations associées à une règle r_i, donne les règles applicables après l'application de r_i. L'ensemble des saturations donne les règles qu'on ne peut plus appliquer après.

Chaque entrée du dictionnaire fait référence à un modèle morphologique qui décrit le comportement linguistique des éléments qui lui font référence. Ces modèles contiennent eux même des ensembles de validations et de saturations qui seront combinés à ceux des règles de grammaire appliquées.

Le dictionnaire est organisé en liste, chaînée par ordre alphabétique des éléments dominants (relation "*frère direct* "). Les éléments dominants sont des éléments qui ne sont préfixes d'aucun autre élément du dictionnaire. Ces éléments dominent des sous-listes contenants leurs différents éléments préfixes appartenant au dictionnaire, classés par longueur décroissante (relation "*fils direct* "). Afin d'optimiser les accès au dictionnaire, la liste chaînée des dominants est décomposée en p sous-listes, accessibles à partir d'un arbre binaire comprenant p dominants. Pour plus de détails sur cette organisation, se référer à [GRAN 76].

Le système d'analyse syntaxique destiné à produire la (ou les) structure(s) possible(s) pour une phrase donnée est constitué de deux modèles :

- Un filtre de dépendances, qui a pour objet de construire toutes les arborescences possibles en utilisant une liste de relations de dépendance entre les catégories. Le principe est le suivant : une fois déterminé un candidat gouverneur, déterminer un sous-arbre gauche et un sous-arbre droit, puis par une technique de retour arrière, construire tous les sous-arbres possibles.

- Un filtre grammatical, constitué par une grammaire "hors contexte", qui permet d'effectuer des contrôles sur les structures (ou sous-structures) produites par le filtre précédent.

Suivant les applications visées, l'analyseur de dépendances seul, ou les deux modèles syntaxiques pourront être utilisés.

Il est à noter que le système P.I.A.F. ne limite ni le nombre ni la nature des filtres mis en oeuvre. Un filtre sémantique a pu être ainsi développé

[JOLO 78] en introduisant la notion de trait sémantique pour distinguer des catégories syntaxiques différentes qui n'ont pas lieu d'exister sur un plan purement grammatical.

Le système P.I.A.F. a été appliqué au domaine de la documentation automatique et a donné lieu aux logiciels PIAFDOC [GRAN 80] et PIAFPS [MERL 82].

Dans le logiciel PIAFDOC, seul le transducteur d'états finis a été utilisé pour réaliser l'analyse morphologique. Le logiciel est employé aussi bien lors de l'indexation des documents que lors de l'interrogation pour analyser les requêtes considérées comme des énoncés en langue naturelle. Le dictionnaire a été enrichi d'informations d'ordre sémantique pour permettre d'effectuer une normalisation des mots-clés. A chaque entrée du dictionnaire ont pu être associés un certain nombre d'indicateurs décrivant son comportement documentaire (mot vide, mot composé, possède un synonyme préférentiel,etc...). Ce mélange d'informations d'origines différentes, grammaticales et sémantiques, s'est révélé difficilement contrôlable par la suite. Mais le problème majeur de l'utilisation de ce logiciel reste la résolution des homographies et des polysémies qui est à la charge de l'utilisateur qui doit les lever de manière interactive. Ce qui provoque de nombreux arrêts de l'analyse et des interventions humaines fastidieuses. C'est la raison pour laquelle des traitements pré-syntaxiques ont été introduits pour concevoir le prototype PIAFPS.

Dans PIAFPS, le choix d'une pré-syntaxe, par rapport à une analyse syntaxique complète, beaucoup plus coûteuse, est délibéré et est justifié par le fait que de nombreuses ambiguïtés dues à la morphologie peuvent être résolues, en français, par un examen du contexte immédiat de la forme ambiguë.

Le principe de cette analyse de surface repose sur l'utilisation d'un ensemble de règles, géré par l'utilisateur, permettant d'effectuer deux types d'actions :

- des regroupements,
- des résolutions d'homographies.

Le formalisme employé pour représenter ces règles est celui des règles de production, bien connu en intelligence artificielle :

SI < condition > ALORS < action >

Malgré cette approche séduisante, le caractère fortement interactif du logiciel, tant au niveau de la gestion des informations présentes dans le dictionnaire, qu'au niveau de la gestion de l'ensemble des règles de regroupements et de résolutions d'homographies (l'utilisateur n'a aucune aide pour assurer la cohérence des règles, ou s'assurer de la validité d'une nouvelle règle) demeure, et constitue comme pour PIAFDOC l'obstacle essentiel à son utilisation.

3.4. Le système du projet SYDO

L'objectif de ce projet est la construction d'un système documentaire automatisé, et la première étape est la conception d'un système d'analyse morpho-syntaxique du français [ROUA 83]. Ce système se compose de quatre modules :

- Un prétraitement morphologique.
- Une analyse morphologique.
- Un prétraitement syntaxique.
- Une analyse syntaxique.

Le prétraitement morphologique comprend la saisie du texte à l'aide d'un éditeur spécialisé GAPRET imposant des conventions de saisie [ANTO 84] et réalisant des modifications du texte d'entrée afin de faciliter l'analyse morphologique proprement dite.

Les conventions de saisie imposent par exemple que chaque forme du texte soit séparée des formes voisines par un blanc. Cette convention permet entre autre, de distinguer le point ponctuation, du point abréviation. Les modifications peuvent être des substitutions des caractères majuscules marquant le début d'une phrase par les caractères minuscules correspondant, des remplacement de formes élidées telles que " *j'* " par leur forme complète " *je* ", ou des éclatements d'amalgames telles que l'article contracté " *aux* " en la préposition " *à* " et le déterminant " *les* ". Lors de cet éclatement certaines formes ambiguës telles que " *les* " dans l'exemple précédent pourront être catégorisées directement de manière non-ambiguë :

" *aux* " --> (" *à* ", P) + (" *les* ", D)

Ceci évite de créer une ambiguïté supplémentaire pour l'analyse morphologique.

L'analyse morphologique [GALI83], basée sur le modèle linguistique défini par A. Berrendonner [BERR 83] réalise une transduction permettant d'associer à chaque forme du texte prétraité une catégorie, ou plusieurs en cas d'homographie, et un ensemble de variables grammaticales. Cette analyse morphologique est classique et, pour une forme du texte, effectue une décomposition en base et flexion, ce découpage devant être validé par un ensemble de règles. le modèle linguistique utilisé comprend 11 catégories :

F : les noms et adjectifs
V : les verbes
D : les déterminants
Y : les particules préverbales
C : les conjonctions de coordination
Q : les conjonctions de subordination
W : les adverbes
T : les ponctuations
P : les prépositions
G : les négations
H : les prophrases " *oui* " , " *si* "

Tout un système de variables permet de sous catégoriser ces classes syntaxiques.

Exemples :

la variable NA (type nominal) associée à la catégorie F, prend ses valeurs dans l'ensemble { NOM, ADJ, NAN } , pour nom, adjectif, non déterminé, Une variable NN (sous-type nominal) affectera les mots marqués F(NOM).

Le résultat de l'analyse morphologique effectuée pour chaque forme indépendamment du contexte, contient des ambiguïtés grammaticales dues aux homographies, même si le petit nombre de catégories utilisées avec ce modèle linguistique en élimine un certain nombre.

Afin de lever les ambiguïtés grammaticales issues de l'analyse morphologique, qui peuvent être résolues par un examen restreint du contexte, un filtre statistique de levée automatique d'ambiguïtés grammaticales a été développé. Ce filtre, construit à partir d'un échantillon de textes du corpus, repose sur le modèle statistique des chaînes de Markov d'ordre variable. Pour de plus amples détails sur l'aspect formel de ce travail, nous renvoyons le lecteur aux travaux de Kallas, [KALL 87], qui montre que le modèle Markovien est un modèle adapté au problème posé, et que l'ordre 2 est

optimal. Cette méthode basée sur les probabilités conditionnelles de successions des catégories grammaticales, calculées sur un échantillon du corpus, propose comme toutes les méthodes statistiques la solution la plus probable, et n'est donc pas à l'abri d'erreurs difficilement récupérables par la suite. Dans le filtre proposé par Kallas, un calcul d'un intervalle de confiance, permet de proposer des alternatives concurrentes lorsque plusieurs solutions ont leur probabilité incluse dans cet intervalle. Ce filtre testé sur un ensemble de cinq résumés comportant 626 mots dont 171 ambigus n'a produit que 3 erreurs (1,7%). Une expérimentation plus poussée devrait permettre de vérifier cette efficacité, car l'incidence de l'échantillonnage réalisé est difficile à mesurer.

Une méthode de levée contextuelle d'ambiguïtés grammaticales reposant sur des critères linguistiques est à l'étude, mais à notre connaissance n'a pas encore été réalisée. Cette méthode est basée sur l'établissement de règles contextuelles du type :

(D , Y) F --> D F

permettant de lever l'ambiguïté déterminant - particule préverbale, lorsque la forme suivante est un nom ou un adjectif.

Dans ce cas le petit nombre de catégories grammaticales utilisées, ne permet pas une analyse morphologique suffisamment fine pour résoudre de manière efficace un grand nombre d'ambiguïtés.

Un certain nombre de traitements effectués à partir des résultats de l'application du filtre statistique pour la levée des ambiguïtés morphologiques, vont précéder l'analyse syntaxique [LALL 86] :

- Délimitation des syntagmes minimaux.
- Traitement des morphèmes discontinus.
- Extraction de syntagmes nominaux.

Un syntagme minimal est un syntagme qui n'en contient pas d'autres. Une variable, notée FF, permet de préciser lors de l'analyse morphologique, en fonction des catégories, et des sous-catégorisations de préciser si une forme est dite "forte", "faible", ou "non-marquée". Un syntagme est considéré comme une suite de formes où les formes faibles se trouvent en tête. En considérant de plus les formes des catégories Q, T et C, dont un des rôles syntaxiques est de borner les syntagmes, un algorithme de découpage a été réalisé.

Le traitement des morphèmes discontinus tels que les négations "*ne...pas* " " *ne ... jamais* ", ou les temps composés des verbes, constitue en fait des régularisations consistant par exemple :

- pour les négations,
  à distribuer la particule " *ne* " sur chacun des corrélats négatifs de la phrase,
  à transformer le corrélat négatif " *jamais* " en " *pas une fois* ",
  puis, à regrouper les particules " *ne* " et " *pas* " en " *ne+pas* ".

- pour les verbes conjugués aux temps composés,
  à ramener le participe passé devant l'auxiliaire :
  " *ils ont mangé* " devient " *ils mangé-ont* "
  et la forme ainsi regroupée sera analysée comme :
  " *mangé-ont* , manger V"

L'extraction des syntagmes nominaux s'opère par segmentations successives du texte. Dans un premier temps une segmentation en propositions est effectuée en utilisant l'algorithme de Maegaard et Spang-Hanssen (cf. précédemment 3.2.). Il faut ensuite repérer dans chaque proposition le verbe fini (verbe à une forme personnelle), qui par son comportement syntaxique doit permettre de repérer les syntagmes nominaux régis (les propriétés rectionnelles des verbes sont contenues dans le lexique). Nous ne savons pas si ce module est opérationnel.

A l'issue de ces traitements, l'analyse syntaxique pour le syntagme nominal développée par G. Lallich-Boidin est effectuée. Cette analyse repose sur la reconnaissance d'indicateurs de structures lexicaux, ISL, qui sont les formes possédant des propriétés rectionnelles (verbes, noms, adjectifs), c'est-à-dire régissant des compléments. Ces informations syntaxiques sont contenues dans le lexique. Les ISL renseignent sur la structure des syntagmes auxquels ils participent. L'algorithme s'appuie sur le modèle hors-contexte et est une adaptation de l'algorithme de Earley [LALL 86].

Actuellement, le principal problème de cet analyseur automatique du français, est l'enchaînement des différents modules que nous venons de voir. En particulier, l'analyse syntaxique présuppose un texte analysé morphologiquement, ne comportant plus ni ambiguïtés, ni erreurs.

De plus, il nous parait utopique de vouloir consigner dans le lexique, des informations syntaxiques comme les propriétés rectionnelles, car d'une part, ces informations doivent être déterminés par des spécialistes, et que d'autre part, les lexiques, surtout dans des applications documentaires, sont sujets à de perpétuelles évolutions (enrichissement du vocabulaire). Malgré les travaux des linguistes ayant entrepris des études systématiques de propriétés syntaxiques visant à l'exhaustivité [GROS 75], cet écueil nous parait

insurmontable, au moins dans l'état actuel des connaissances linguistiques pour tout système voulant conserver un certain caractère de généralité.

3.5. Un système basé sur des algorithmes à apprentissage

L'équipe constituée autour de A. ANDREEWSKY a développé un système d'analyse automatique de textes en langue naturelle ayant donné lieu à de nombreuses publications et thèses parmi lesquelles : [ANDR 73 et 77], [FLUH 77], [HLAL 77], [DEBI 77 et 82]. Ces traitements linguistiques ont été orientés vers la réalisation de systèmes documentaires et ont abouti à la réalisation du logiciel SPIRIT, [FLUH 81 et 85].

L'idée centrale de ce système est l'acquisition automatique d'informations linguistiques, et ce à tous les niveaux d'analyse abordés, morphologique, syntaxique et sémantique, pour des univers textuels ouverts.

L'analyse grammaticale (analyse morphologique et analyse syntaxique de niveau grammatical) réalisée par C. Fluhr est fondée sur l'acquisition de relations positionnelles entre les mots de la langue. Une relation positionnelle est une relation syntaxique traduisant la possibilité ou l'impossibilité de succession (ou de proximité géographique) de deux classes de mots de la langue :

Exemples :

relation entre l'adjectif épithète et le substantif,
relation verbe-adverbe,
article-substantif,
etc...

L'hypothèse ayant servi de base à la méthode, est que les relations positionnelles permettent de lever la plupart des ambiguïtés grammaticales qu'il est important de résoudre dans le cadre d'une application documentaire. Cette méthode devant notamment convenir pour le français et l'anglais qui sont des langues très positionnelles et peu morphologiques.

L'analyse grammaticale est basée sur des algorithmes à apprentissage. C. Fluhr en donne cette définition :

" *Intuitivement, et par analogie avec la notion d'apprentissage chez l'homme, on peut appeler algorithme à apprentissage un algorithme dont les réponses s'améliorent en moyenne au cours du temps (en fait par rapport à un but ou une sanction déterminée)* ."

La phase d'initialisation de l'apprentissage consiste, à partir d'un échantillon de textes résolus manuellement (attribution manuelle des valeurs grammaticales aux mots du texte) à coder de nouveaux textes dont les erreurs sont corrigées par un "professeur" et sur lesquels s'effectue l'apprentissage. Cet apprentissage a consisté à élaborer quatre modules d'analyse grammaticale :

-1) Une analyse basée sur l'acquisition de règles binaires fréquentielles :

Exemple d'une règle non-ambiguë :

(PROS) * (ARTD, PROV) --> PROS * PROV

(" *je les ...* ")

Exemple d'une règle ambiguë :

(ARTD, PROV) * (SUBS, VT) --> ARTD * SUBS [f_1]

--> PROV * VT [f_2]

(" *... les portes ...* ")

f_1 et f_2 sont les fréquences calculées des couples solutions, par rapport au nombre d'occurrences de l'ambiguïté :

[f_1] = fréquence ((ARTD*SUBC) / ((ARTD,PROV)*(SUBS,VT)))

et [f_2] = fréquence ((PROV*VT) / ((ARTD,PROV)*(SUBS,VT)))

-2) Une analyse basée sur l'acquisition de règles ternaires fréquentielles :

(PROS) * (ARTD, PROV) * (SUBS, VT) --> PROS * PROV * VT

(" tu les portes ")

(ARTD, PROV) * (SUBS, VT) * (PREP, PREV)

--> ARTD * SUBS * PREP [f1]

--> PROV * VT * PREV [f2]

(" *... les portes à ...* ")

-3) Une analyse basée sur la construction d'une matrice binaire de précédence fréquentielle MAT, où :

$$MAT(V_i, V_j) = \frac{n(V_i * V_j)}{n \text{ couples}}$$

c'est-à-dire que chaque élément de la matrice contient la fréquence du couple de catégories (V_i, V_j) par rapport à l'ensemble des couples.

-4) Une analyse basée sur la construction d'une matrice ternaire de précédence fréquentielle où, par analogie, chaque élément de la matrice contient la fréquence du triplet de catégories par rapport à l'ensemble des triplets.

Les analyses grammaticales d'ordre 2 (règles syntaxiques et matrice de précédence binaires) nécessitent un apprentissage beaucoup moins long que les analyses grammaticales d'ordre 3 (ternaires), et donnent dans un premier temps de meilleurs résultats. C'est l'analyse grammaticale basée sur la matrice de précédence fréquentielle binaire qui "converge" (apprentissage pratiquement terminé) le plus vite.

On peut rappeler ici que les matrices de précédence sont des outils bien connus dans le domaine de la compilation des langages de programmation [COLM 70].

C. Fluhr a obtenu à partir d'un texte de 3000 mots, résolu morphologiquement, une syntaxe permettant de traiter avec 93% de réussite un texte complètement nouveau, ce qui permet d'affirmer que les hypothèses qui sont à la base de la méthode sont vérifiées. Cette méthode d'analyse a également été testée avec succès pour des textes en anglais. Les résultats obtenus pour le russe et l'arabe sont encourageants [HLAL 77].

F. Debili a développé une analyse syntaxique basée sur l'évaluation de chemins à partir de matrices de précédence fréquentielles binaires et ternaires [DEBI 77]. Cette méthode consiste à évaluer à partir des poids des chemins élémentaires, les poids des chemins complets concurrents.

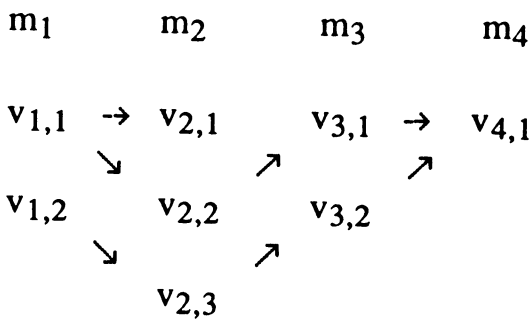
Un chemin élémentaire binaire entre deux formes consécutives (dans un texte) m_i et m_{i+1} est donné par un ensemble de deux catégories grammaticales (v_i * v_{i+1}) telles que : v_i soit une catégorie grammaticale de m_i et v_{i+1} soit une catégorie grammaticale de m_{i+1}, et telles que $M(v_i, v_{i+1})$

l'élément de la matrice de précédence associée, ait une valeur non nulle (possibilité de succession).

Un chemin complet binaire pour un texte T de n formes, est constitué par un ensemble de n-1 chemins élémentaires binaires permettant de passer de la première forme de T à la dernière, et telles que la catégorie grammaticale extrémité de chaque chemin élémentaire binaire (sauf pour le dernier) soit la catégorie grammaticale origine du chemin élémentaire binaire suivant.

Exemple :

sur cet exemple les chemins élémentaires binaires sont matérialisés par des flèches :



($v_{1,1} * v_{2,2}$) et ($v_{1,2} * v_{2,3}$) sont 2 chemins élémentaires binaires.

($v_{1,1} * v_{2,2} * v_{3,1} * v_{4,1}$) et ($v_{1,2} * v_{2,3} * v_{3,2} * v_{4,1}$) sont les 2 seuls chemins complets binaires pour le texte T.

Le poids associé à un chemin complet binaire est obtenu en faisant la moyenne (soit arithmétique, soit géométrique) des poids associés aux chemins élémentaires binaires qui le composent. Le chemin complet de plus fort poids pour un texte T ambigu (qui contient au moins un couple de formes consécutives, auquel est associé plusieurs chemins élémentaires binaires), est fourni comme solution de l'analyse syntaxique binaire.

L'analyse syntaxique ternaire est basée sur l'évaluation de chemins complets ternaires à partir des poids associés aux chemins élémentaires ternaires. La méthode est identique à celle de l'analyse syntaxique binaire que nous venons de détailler.

F. Debili a développé ensuite, un module d'analyse syntaxico-sémantique dans le but de reconnaître et d'extraire automatiquement un certain nombre de connexions qui sont établies entre les mots de la langue (C. Fluhr à la suite de son analyse grammaticale, avait esquissé ce type de traitement dans le cadre

d'une application à un système documentaire, notamment les relations de synonymie à l'intérieur d'un corpus, à partir du contexte sémantique du concept dans le corpus, appelé champ sémantique [FLUH 77]).

L'objectif de ce module est de reconnaître deux types de relations lexicales sémantiques (RLS), dont nous donnons les définitions et des exemples, extraits de [DEBI 82] :

-1) RLS syntagmatiques.

Une relation lexicale sémantique syntagmatique est : "*une relation définie sur un n-uplet de mots pleins extraits d'un texte et vérifiant certaines conditions morpho-syntaxiques* ".

C'est à partir de la composante syntaxique du système qui permet de construire une représentation dépendancielle des relations potentielles (analogue aux "stemmas" de L. Tesnière [TESN 59]), que seront, lors de l'application de la composante sémantique en utilisant les RLS paradigmatiques, déterminées ces relations syntagmatiques, dont nous donnons quelques exemples :

- La relation Substantif-adjectif Antérieur (SA) : "*un maigre repas* "
- La relation Substantif-Substantif (SSp) : " l*a porte de la maison* "
- La relation Verbe-complément Indirect (VI_p) : " *rayé de la liste* "

F. Debili a défini en tout neuf RLS syntagmatiques qu'il s'est donné pour objectif de reconnaître.

-2) RLS paradigmatiques.

Une relation lexicale sémantique paradigmatique est : " *une relation qui s'établit entre les unités lexicales qui sont susceptibles de commuter* ".

Ces relations paradigmatiques sont déterminées, de manière semi-automatique, à partir d'une description paradigmatique des éléments de la phrase. Cette description paradigmatique résulte de la composante morphologique du système, qui est fondée sur la détermination automatique de familles de mots, qui seront représentées par des éléments issus de ces familles, les "représentants".

Cette méthode de détermination de familles est basée sur le principe du découpage d'un mot en préfixe, radical, suffixe, désinence, qui permet d'inférer une proximité morphologique. La validité du découpage étant vérifiée par des heuristiques basées sur la compatibilité des préfixes ou suffixes.

Voici un exemple de famille de mots :
F(ACH) = (ACHAT, ACHETER, ACHETEUR, RACHAT, RACHETER)

C'est à partir de ces familles de mots que sont construites les RLS paradigmatiques basées sur une proximité morphologique, dont voici quelques exemples :

- La relation " nom d'action " ACT :
  ACT(*afficher*) = *affichage*, ACT(*interdire*) = *interdiction*

- La relation " nom d'agent " AG :
  AG(*acheter*) = *acheteur*

- La relation " d'antonymie " ANT :
  ANT(*responsable*) = *irresponsable*
- La relation de " synonymie " SYN :
  SYN(*habitat*) = *habitation*

La nature de ces relations est inspirée des travaux de A.K. Zholkolvskij et I.A. Mel'tchuk sur les "fonctions lexicales" [ZHOL 67].

F. Debili fait remarquer que les informations que sont les familles de mots sont à la base des substitutions lexicales dans la formation des paraphrases du type :

" *Interdiction d'afficher sur les murs.* "
" *Affichage mural interdit.* "
" *l'affichage sur les murs est interdit* "
etc...

Pratiquement, les résultats donnés sur la constitution automatique de familles suffixales (sans le traitement analogue des préfixes) engendre un pourcentage de familles incorrectes d'environ 5%, qui ont subi une correction manuelle.

Les performances globales de l'analyse syntaxico-sémantique, mesurées sur un texte de 4700 mots comprenant 1011 RLS syntagmatiques, ont donné

89% de reconnaissance de relations correctes, 3,3% de reconnaissance de relations erronées, et 7,7% de non-reconnaissance.

Le cadre même des domaines d'application envisagés au sein de cette équipe, les systèmes de recherche d'informations, autorise l'utilisation de méthodes produisant un minimum d'erreurs, ce qui est le cas de tous les traitements statistiques, qui ne produisent pas toujours des résultats absolus.

Cette analyse du français fondée sur des algorithmes à apprentissage, a été appliquée à d'autres domaines que celui de la recherche d'informations, notamment au domaine de la reconnaissance automatique de la parole.

Le modèle linguistique défini dans ce système utilise un grand nombre de catégories grammaticales (plus de 150), et en conséquence les outils réalisés, notamment les matrices de précédence d'ordre 3, sont énormes et nécessitent des formalismes de représentation appropriés (techniques de compression des matrices creuses).

Sur le plan méthodologique, le refus délibéré de faire appel à des connaissances sémantiques prédéfinies du type des traits sémantiques [KATZ 64], comme le sont toutes les sous-catégorisations, nous parait être essentiel pour tout système dont la vocation est d'appréhender des textes ouverts (sans restrictions sur les domaines).

3.6. Conclusion

Le tour d'horizon que nous venons d'effectuer ne prétend pas être exhaustif, et l'on ne préjuge en aucun cas de l'intérêt des systèmes d'analyse qui ne sont pas présentés dans ce chapitre.

Si le problème de l'analyse morphologique automatique de textes en langues naturelles (français, anglais, etc...) est généralement bien résolu, la persistance de certaines ambiguïtés grammaticales, dont la levée relève de critères syntaxiques, voire sémantiques est un frein à toute tentative d'analyse syntaxique automatique de la langue naturelle. Ces ambiguïtés, inhérentes au langage naturel, compliquent et augmentent considérablement les coûts d'exécution dès que l'on tente de les lever automatiquement. Les solutions généralement employées consistent à restreindre les champs d'application (traitement d'un sous-langage du langage naturel) de manière à permettre de déterminer a priori un certain nombre de propriétés syntaxiques ou

sémantiques (la limite n'étant d'ailleurs pas toujours facile à établir, dans la mesure où il en existe une...), utiles à la résolution de ces ambiguïtés.

Dans ce sens, toutes les tentatives visant à extraire automatiquement ce type de renseignements linguistiques nous paraissent très intéressantes, tout comme le sont les constitutions de lexiques exhaustifs par rapport à ces types de propriétés, sachant bien dans ce dernier cas que les langues naturelles sont des langues vivantes en perpétuelle évolution : "*des spécialistes ont estimé que, chaque année, au moins 4000 expressions ou termes nouveaux viennent s'agréger au vocabulaire français* " extrait du "bon usage" de M. Grevisse [GREV 80].

4. Conclusion

L'objectif de notre travail est la conception d'un analyseur du français destiné à être intégré dans le processus d'indexation d'un système de recherche d'informations entièrement automatisé. Le but visé est l'extraction automatique, à partir d'un texte écrit en langue naturelle, d'éléments structurés contenant l'essentiel de la connaissance véhiculée (ce sont les Groupes Conceptuels que nous définissons au chapitre suivant)

Un tel analyseur devant permettre de traiter n'importe quel texte écrit, en quantité importante et de manière complètement automatique, nous avons dégagé deux priorités essentielles en vue de réaliser des outils bien adaptés au contexte de l'application :

- définir des outils d'analyse performants,
- définir des outils d'analyse capables d'appréhender des univers textuels ouverts.

-1)Définition d'outils d'analyse performants

Nous opérons dans le cadre d'une base de données textuelles, c'est-à-dire que nous disposons du texte plein, dans son intégralité. Il nous faut donc définir des outils d'analyse aptes à traiter des corpus importants, d'où notre orientation vers une analyse partielle de la langue naturelle.

Notre objectif étant la reconnaissance des groupes conceptuels, il n'est pas utile de disposer de la structure (ou des différentes possibilités de structures) syntaxique complète de chaque phrase, ce qui nécessiterait d'ailleurs des mécanismes d'analyse complexes et coûteux.

Nous avons renoncé à utiliser les techniques descendantes ("top-down") classiques en analyse automatique des langues naturelles, qui nécessitent un niveau de description conséquent de la langue analysée. Ce choix résulte du fait qu'il n'existe pas de description complète du français, et que les grammaires écrites à ce jour se rapportent à des sous-langages plus ou moins importants du langage naturel. De plus ces grammaires utilisent des renseignements linguistiques de type "syntaxico-sémantico-pragmatique", que nous ne prenons pas en compte, car difficiles à généraliser systématiquement, comme par exemple certaines contraintes d'ordre sémantique énoncées dans l'analyse en chaîne de M. Salkoff, ou comme la sous-catégorisation d'ordre syntaxico-sémantique utilisée dans l'analyseur du projet SYDO. Nous avons mentionné cet état de fait, lors de la présentation de ces analyseurs.

D'autre part, ces techniques d'analyse produisent généralement, en utilisant la technique coûteuse du retour arrière, un ensemble de solutions parmi lesquelles se trouve la ou les bonnes interprétations, ensemble qu'il s'agit de filtrer par la suite pour éliminer les solutions incorrectes (ce filtrage étant obtenu par l'exploitation de renseignements de type "sémantico-pragmatique").

Nous nous sommes donc orientés vers une analyse ascendante, c'est-à-dire gouvernée par les données, qui ne peut être que partielle, car nous utilisons uniquement des renseignements linguistiques fournis par la morphologie (qui est universelle) et par un ensemble de relations positionnelles ne portant que sur la structure de surface des textes analysés.

-2) Appréhension d'univers textuels ouverts

Dans le but d'assurer une certaine généralité au système, nous avons renoncé à reprendre la démarche classique consistant à définir a priori un environnement sémantique des concepts. Nous n'établissons pas, non plus, de listes de mots vides (sans information sémantique); par contre, nous considérons que certaines catégories grammaticales (articles, pronoms, etc...) ne représentent que des outils, utiles pour la levée des ambiguïtés grammaticales (homographies).

Ces ambiguïtés grammaticales inhérentes au langage naturel constituent le principal obstacle à tout traitement automatique de textes en langue naturelle. Une des caractéristiques essentielles de notre analyse est de ne tenter la levée

de ces ambiguïtés que dans le cas où cela est utile (c'est-à-dire pour l'extraction des groupes conceptuels).

Afin d'obtenir une analyse entièrement automatique, il est nécessaire d'effectuer un traitement particulier pour les mots inconnus, dont la rencontre est inévitable en cours d'analyse dès que l'on appréhende des textes "ouverts" (dont le domaine n'est pas fermé, et donc dont le vocabulaire n'est pas entièrement connu a priori). Ce traitement doit permettre une interprétation linguistique des mots inconnus, et éventuellement un enrichissement automatique du vocabulaire, afin d'éviter que le processus d'analyse ne s'arrête dès qu'il ne possède pas les informations nécessaires. Pour ce faire, il est important que les informations linguistiques que l'on peut associer aux mots de la langue (en fonction d'un modèle linguistique déterminé) puissent être retrouvées automatiquement (ou éventuellement approchées) dans le cas des mots inconnus, à partir d'un ensemble de constantes linguistiques disponibles dans le système d'analyse.

Ce sont ces deux priorités qui nous ont conduit à développer une stratégie d'analyse de surface pour réaliser nos objectifs. Nous ne reviendrons pas, dans la présentation de ce travail, sur les notions "Chomskiennes" très classiques de "structure profonde" et de "structure de surface" du langage [CHOM 57 et 65]. Nous précisons simplement que l'analyse linguistique que nous effectuons est une analyse morpho-syntaxique partielle, se situant au niveau de la structure de surface de la langue.

L'hypothèse sur laquelle se fonde notre analyse est qu'il n'est pas nécessaire d'effectuer une analyse morpho-syntaxique complète d'un texte pour reconnaître certaines structures particulières : on peut résoudre une grande partie des ambiguïtés grammaticales en se contentant d'une analyse syntaxique de surface. Cette hypothèse ne peut être vérifiée que par les résultats expérimentaux. Certaines expériences menées dans le même sens ont donné des résultats suffisamment satisfaisants en ce qui concerne le français, pour nous conforter dans notre démarche [FLUH 77], [MAEG 78], [DEBI 82] et [MERL 82].

Nous entendons par analyse syntaxique de surface, une analyse de la syntaxe primaire de la langue naturelle : c'est-à-dire la reconnaissance de la structure des syntagmes principaux. Une phrase pourra être analysée dans sa continuité lorsque sa syntaxe est suffisamment simple pour le permettre (ex : SN V SN), ou pourra être reconnue par morceaux lorsque la complexité, inhérente à la langue naturelle (propositions relatives fréquentes par exemple), nécessiterait l'utilisation d'un processus d'analyse beaucoup plus compliqué et plus coûteux.

L'utilisation de règles de grammaire simplifiées, telles que les règles positionnelles [ANDR 73] et [FLUH 77], qui vérifient la pertinence de la succession dans une phrase de deux catégories grammaticales, et telles que les règles de regroupement et de simplification [MERL 82], permettant la reconnaissance de locutions ou de constituants syntagmatiques particuliers, nous paraissent des outils particulièrement bien adaptés à cette approche.

La stratégie utilisée dans le processus de segmentation automatique du français [MAEG 78], qui consiste à reconnaître un petit nombre de marqueurs syntaxiques jouant le rôle d'indicateurs de structures, au cours d'une analyse de surface, puis à traiter ce résultat par un automate déterministe permettant de réaliser l'application désirée, correspond tout à fait à la démarche que nous mettons en oeuvre.

Nous rappelons que notre objectif n'est pas une analyse en vue d'une compréhension de la langue naturelle, mais une reconnaissance de certains syntagmes susceptibles de véhiculer le sens contenu dans un texte, et d'un nombre restreint de relations prédéfinies entre ces structures.

Pour cela, nous verrons au chapitre suivant (cf. III.3) que l'on doit obtenir au moyen de l'analyseur de surface une structure linéaire (chaîne de solutions morphologiques) traitable par un transducteur d'états finis déterministe destiné à extraire les groupes conceptuels. A ce propos, le rôle syntaxique prépondérant joué par certaines catégories grammaticales "outils" : tels que les déterminants (articles, adjectifs déterminatifs), prépositions, conjonctions, pronoms, etc..., s'apparente à des "indicateurs de structures syntagmatiques" permettant de délimiter et de repérer des groupements syntagmatiques (nominaux, verbaux, adjectivaux, etc...). Nous nous sommes donc plus particulièrement intéressés à leur "comportement". En effet la liste des éléments de ces catégories grammaticales est relativement restreinte, et ils sont facilement identifiables dans les textes au cours de la phase d'analyse morphologique. Nous reviendrons sur ce point essentiel lors de la définition du modèle linguistique (cf. IV.2).

Nous allons maintenant présenter dans le chapitre III, la définition des Groupes Conceptuels, et le processus permettant de les extraire à partir d'une analyse de surface des textes. Nous détaillerons ensuite dans les chapitres IV et V, les moyens que nous nous donnons pour réaliser cet objectif, en détaillant le modèle linguistique utilisé, le formalisme des données manipulées, et l'analyseur de surface de la langue naturelle réalisé.

CHAPITRE III

PLAN DU CHAPITRE III

LES GROUPES CONCEPTUELS

LES GROUPES CONCEPTUELS

1. Introduction

Un des principaux facteurs qualitatifs des SRI repose sur la définition des termes d'indexation, et particulièrement sur leur potentialité de signification. Le contenu sémantique d'un texte est exprimé par les éléments véhiculant une information sémantique d'une part, et par leur arrangement, leur structuration syntaxique d'autre part, qui définit des relations entre ces concepts, et par là des constructions sémantiques complexes. De ce fait, plus le degré potentiel de structuration des termes d'indexation sera proche de celui de la langue naturelle (donc élevé), et mieux l'adéquation entre le contenu sémantique d'un texte et son indexation pourra être réalisée. Cela à condition que les constituants de ces termes d'indexation comprennent les éléments les plus informatifs sémantiquement de la langue.

La nécessité de tenir compte de la structuration syntaxique de portions de texte (syntagmes, propositions, phrases, etc...) dans la représentation des termes d'indexation (et même pour une représentation de la connaissance dans d'autres applications), s'est avérée à la suite du constat des limitations liées à l'utilisation de l'indexation par mots simples. La non reconnaissance des mots

composés, et la quasi impossibilité de les retrouver à partir de mots simples, a constitué un des éléments décisifs en faveur de l'introduction d'une structuration issue de la syntaxe.

L'intérêt des Groupes Conceptuels, tels qu'ils sont définis dans [KERK 84], est de représenter sans déformation les portions de texte susceptibles de contenir des concepts de plus haut niveau, qui reflètent un thème (ce dont il est question). Cette représentation est normalisée, afin de pouvoir en extraire les termes d'indexation.

Il est nécessaire de réaliser un compromis entre le coût de l'analyse (au sens développement et application), et le niveau de structuration souhaité, car plus ce degré potentiel de structuration augmente, et plus le processus de reconnaissance devient complexe, et par conséquent coûteux sur de grands volumes d'information. La recherche d'un tel compromis est également dictée par l'état de l'art en matière d'analyse et de compréhension de la langue naturelle : il n'existe pas encore de modèles susceptibles de conduire à des processus non ambigus.

Dans un premier temps, nous nous limitons à une reconnaissance de portions de texte à l'intérieur d'une même phrase : tous les éléments d'un groupe conceptuel participeront de la même phrase, et nous ne considèrerons comme contexte d'un groupe conceptuel que les éléments de la phrase englobante. Cette limitation à la phrase peut se comprendre dans une application visant l'indexation de textes en langue naturelle, il en irait autrement pour des applications visant une compréhension de ces textes, où il serait alors nécessaire de prendre en compte les relations supraphrasiques.

Nous commencerons dans ce chapitre par définir la notion de Groupes Conceptuels, en exposant les choix effectués pour déterminer les éléments qui les composent, et en en donnant une syntaxe cible, c'est-à-dire une syntaxe de représentation normalisée. Nous indiquerons ensuite les possibilités d'interconnecter les groupes conceptuels issus d'une même phrase. Puis nous présenterons une méthode simple, basée sur la construction d'un transducteur d'états finis déterministe, d'extraction automatique de ces groupes à partir d'une analyse de surface d'un texte.

2. Les groupes conceptuels

De nombreuses études ont été menées en linguistique générale ou appliquée, qui démontrent une relative concordance entre syntagmes nominaux et thèmes d'une part, et syntagmes verbaux et rhèmes d'autre part. Dans les systèmes de recherche d'informations possédant une analyse textuelle automatique, on peut remarquer qu'il en est de même, et que l'analyse des syntagmes nominaux tient une place prépondérante. Nous en avons vu quelques exemples au chapitre précédent [SALK 73], [DEBI 82], [FLUH 85], [LALL 86].

Nous citerons comme référence la synthèse de Le Guern [GUER 82] sur le sujet, qui précise que :

" *Si l'on reprend le schéma classique, qui décompose chaque maillon de la communication informative en deux éléments, le thème, indiquant de quoi on parle, et le rhème, ce qu'on en dit, on peut admettre que l'interrogation d'une base de données consiste à poser un thème afin de recueillir les rhèmes qui sont associés à ce thème dans le corpus.* "

C'est donc logiquement les syntagmes nominaux qui nous ont plus particulièrement intéressé pour la définition des GC

2.1. Choix des constituants des Groupes Conceptuels

Le choix des constituants retenus pour les GC repose sur quelques considérations linguistiques que nous développons ci-dessous. Nous pouvons grossièrement répartir les mots (au sens classique du terme) d'une langue naturelle en deux classes :

-1) l'une comprenant les mots-outils, c'est-à-dire des éléments pratiquement vides de sens en eux-mêmes (ils sont souvent appelés mots vides), et dont le rôle essentiel est de déterminer la structure syntaxique (déterminants, pronoms, conjonctions...) : ils jouent le rôle de constructeurs ou "d'articulateurs" (connecteurs pragmatiques) dans la phrase.

-2) l'autre comprenant les éléments véhiculant du sens, c'est-à-dire une fraction plus ou moins riche du thème contenu dans la portion de texte considéré. Cette classe regroupe principalement les substantifs, les verbes, et leurs qualifiants (adjectifs qualificatifs, adverbes...).

Cette classe d'éléments significatifs regroupe en fait, l'ensemble des constituants significatifs de deux structures bien connues de la langue naturelle : les syntagmes (ou groupes) nominaux et verbaux. Parmi les constituants des syntagmes nominaux nous avons retenu les plus "informatifs", dont les classes syntaxiques sont :

- les substantifs,
- les adjectifs qualificatifs et numéraux,
- les participes passés,
- les verbes à l'infinitif,
- ainsi que certaines prépositions.

Nous allons maintenant justifier le choix de ces constituants, en exposant le rôle classique joué par les éléments de ces différentes catégories grammaticales informatives dans la langue.

1) Le substantif (simple ou composé) :

" *est le mot qui sert à désigner, ' à nommer' les êtres animés et les choses; parmi ces dernières, on range, en grammaire, non seulement les objets, mais encore les actions, les sentiments, les qualités, les idées, les abstractions, les phénomènes, etc...*" .

Cette définition se trouve dans "le bon usage" de M. Grevisse [GREV 80].

2)Traditionnellement, les adjectifs sont répartis en deux classes :

-i) les adjectifs qualificatifs dont le rôle est de qualifier et de déterminer (au sens de distinguer une espèce particulière) le nom qu'ils accompagnent.
La forme adjectivale du participe présent est considérée comme un adjectif qualificatif, dans notre classification.

-ii) les adjectifs déterminatifs dont la fonction est d'introduire les noms qu'ils précèdent dans le discours. Ce sont les adjectifs numéraux, possessifs, démonstratifs, relatifs, interrogatifs ou exclamatifs, et indéfinis.

Parmi les adjectifs non qualificatifs, nous avons décidé de retenir les numéraux (aussi bien cardinaux, "deux, trois, ...", que les ordinaux "premier, deuxième, ...") car leur signification est intrinsèque : ils n'ont généralement pas de référence anaphorique.

3) Les participes passés sont considérés au même titre que les adjectifs qualificatifs. Lorsqu'il est épithète, on a alors affaire sans ambiguïté à la forme adjectivale du participe passé. En position d'attribut, la distinction entre forme verbale et forme adjectivale nécessite des considérations d'ordre syntaxique ou sémantique, comme par exemple dans les expressions suivantes :

- dans " *La voiture est empruntée* " on a la forme adjectivale,

- dans " *La voiture est partie* " on a la forme verbale, (*partir* se conjuguant avec l'auxiliaire *être*).

Notre système d'analyse linguistique ne disposant de connaissances ni sémantiques, ni syntaxiques de nature à nous permettre d'effectuer ce type de distinctions, nous avons décidé arbitrairement (dans un premier temps), de prendre en compte systématiquement le participe passé, lorsqu'il se trouve en position d'attribut. Même si dans le cas de la forme verbale la qualification du sujet par le participe passé est moins significative que pour la forme adjectivale en position d'épithète, il s'agit bien tout de même d'une relation de qualification.

Ainsi des deux expressions précédentes, nous déduirons, conformément à la syntaxe de représentation normalisée que nous donnons plus loin (cf. II.2.2), les deux groupes conceptuels suivants :

voiture (*emprunté*) et *voiture* (*parti*)

4) En ce qui concerne les verbes à l'infinitif, le choix de les conserver peut se justifier simplement par l'exemple de l'expression suivante :

" *maison à vendre* "

où l'infinitif, précédé de la préposition " *à* ", est un complément de nom au même titre que dans l'expression :

" *maison à la vente* "

Ces deux expressions nous fourniront les deux groupes conceptuels suivants :

maison ***à*** *vendre* et *maison* ***à_det*** *vente*

où ***à_det*** signifie que la préposition ***à*** est suivie d'un déterminant (ici l'article défini *la.*). Nous reviendrons sur cette particularité, plus avant dans cette section.

On peut rencontrer de la même manière des verbes à l'infinitif participant d'un groupe adjectival comme dans l'exemple suivant :

" *La maison est difficile à vendre* "

qui donnera le groupe conceptuel suivant :

maison (*difficile* ***à*** *vendre*)

M. Grevisse, dans "le bon usage", donne la définition suivante de l'infinitif :

" *L'infinitif est la forme nominale du verbe : c'est proprement un nom d'action; il exprime simplement sans acceptation de personne ni de nombre, l'idée marquée par le verbe* ."

et il précise :

" *Outre qu'il s'emploie dans certains cas avec la valeur purement verbale, l'infinitif remplit, dans bien des emplois, les différentes fonctions du nom* ."

5) Nous retenons également certaines prépositions, parce qu'elles jouent un rôle d'introducteur pour un élément, qu'elles relient à un autre élément de la phrase en pouvant exprimer un certain rapport sémantique :

Exemple :

" *dans* " introduisant un rapport de lieu,
" *pendant* " un rapport de temps, etc...

Elles traduisent explicitement des relations de qualification (nominales ou adjectivales).

Elles jouent également un rôle particulier suivant qu'elles sont ou non suivies d'un déterminant, dans le processus de sélection des termes d'indexation à partir des groupes conceptuels, lorsque ces derniers seront constitués de compléments de noms successifs :

" *l'école de filles de la commune* ".

De cette expression on pourra déduire le groupe conceptuel suivant :

(*école* ***de*** *fille*) ***de***_**det** *commune*

Le lien intime entre la préposition et le nom qu'elle introduit sans article [GREV 80], est utilisé comme critère "syntaxique" dans le processus de cassure des groupes conceptuels utilisé lors de la sélection des termes d'indexation réalisée par le module d'indexation automatique présenté au chapitre I (cf. I.2.1.1.2).

6) Les adverbes font partie de la classe des éléments véhiculant du sens, mais nous ne les avons pas retenus parmi les constituants possibles des groupes conceptuels, car le rôle essentiel d'un adverbe, est de modifier le sens du verbe auquel il est joint. C'est en quelque sorte l'adjectif du verbe, comme l'indique l'étymologie.

Or pour la définition des groupes conceptuels, nous avons décidé de nous restreindre aux groupes nominaux.

Bien que l'adverbe puisse également modifier le sens d'un adjectif comme dans l'exemple suivant :

" *Un livre très bon* "

nous avons considéré par simplification, que l'information supplémentaire dans ce cas, était marginale.

La syntaxe des groupes conceptuels que nous allons présenter étant facilement extensible, l'ajout des adverbes modificateurs d'adjectifs serait, néanmoins, tout à fait possible.

La liste des constituants issus des syntagmes nominaux, et pouvant intervenir dans la composition des groupes conceptuels, étant établie, nous allons présenter une syntaxe cible, destinée à en donner une représentation normalisée.

2.2. Définition de la syntaxe des Groupes Conceptuels

Pour la définition de la syntaxe des Groupes Conceptuels nous avons dû établir un compromis entre la nécessité d'obtenir une certaine complexité de structure pour les GC extraits du texte, et le souci d'efficacité pour un système destiné à traiter de grandes quantités de textes en langue naturelle.

Ces considérations nous ont amené à définir une syntaxe des GC, en considérant ces derniers comme une restriction des syntagmes nominaux, c'est-à-dire non seulement réduits aux constituants significatifs, comme exposé précédemment, mais encore ne contenant qu'un certain nombre des possibilités de relations structurantes des syntagmes nominaux qui sont principalement :

-a) les relations non explicitées de qualification

Exemple :

la relation épithète entre substantif et adjectif, qui se traduit par une juxtaposition de l'adjectif et du substantif :

" *un homme beau* " ou " *un bel homme* "

Nous ne distinguons pas les adjectifs antérieurs des adjectifs postérieurs, car il n'existe pas de règle impérative sur la place de l'adjectif épithète; les tournures stylistiques fréquentes en français les modifient sans cesse. Les adjectifs qualificatifs forment une classe ouverte du français, et il ne serait pas possible en cours d'analyse, lors de la rencontre d'un adjectif qualificatif d'être certain de son caractère antérieur ou postérieur (sauf pour un petit nombre qu'il faudrait connaître a priori) :

Exemple :

" *un avenir proche* " ou " *un proche avenir* "

Par contre la position antérieure ou postérieure d'un adjectif qualificatif sera utilisée lors de l'extraction des groupes conceptuels, principalement pour les syntagmes structurés par des relations de qualifications explicitées par des prépositions.

-b) les relations de qualification explicitées par des prépositions

Exemple :

les compléments de nom introduits par les prépositions :

" *une maison de campagne* " " *une maison en campagne* "

Nous distinguons les compléments introduits par la préposition " *de* " (et sa forme élidée " *d'* "), les articles contractés " *du* " et " *des* " (formes contractées de " *de le* " et de " *de les* ") des compléments introduits par les autres prépositions. Le rapport exprimé par la préposition " *de* " est sémantiquement difficile à cerner, mais le lien établi entre les éléments reliés est beaucoup plus intime.

Certains grammairiens appellent d'ailleurs ce type de prépositions, prépositions vides, par opposition aux prépositions pleines, exprimant un rapport précis. Cette distinction est d'ailleurs utilisée dans le processus de cassure syntaxique des GC lors de la phase de sélection du module d'indexation automatique, tout comme la présence d'un déterminant (cf. 2.1.).

Nous avons déterminé les structures syntaxiques des Groupes Conceptuels que nous souhaitons reconnaître en établissant la syntaxe-cible suivante. Cette syntaxe définit en quelque sorte un squelette normalisé de syntagme nominal (on rappelle qu'il ne s'agit pas de définir une grammaire de reconnaissance, mais simplement un modèle de représentation normalisée).

Tout GC est donc un mot du langage défini par la grammaire donnée ci-après, dans laquelle nous avons fait apparaître en gras la nature des relations structurant les groupes conceptuels.

Dans cette syntaxe cible, les caractères parenthèses "(" et ")" sont des symboles terminaux de la grammaire. Elles permettent de préciser la portée des relations de qualification et expriment donc une relation de dépendance.

Les terminaux , en minuscule, sont normalisés :
- forme au masculin et au singulier pour les adjectifs et les participes passés,
- forme au singulier pour les substantifs.

Les articles contractés sont considérés au même titre que les prépositions; "*du* " et "*des* " donnent "***de*_det** ", et "*au* " et "*aux* " donnent "***à*_*det*** ".

Les non-terminaux sont représentés par des sigles en majuscule dont la signification est donnée en annexe (cf. annexe 3).

GC --> / ARG / NQ / NOM **EP** (ADJ*) **ND** ARG / NOM **EP** (ADJ*) **D** CPN / INF **PREP** CPV
CPV --> / CPN / NOM **EP** (ADJ*)
CPN --> / ARG / CP **PREP** ARG
CP --> / NOM **EP** (ADJ*) / CP **D** NOM **EP** (ADJ*)
NQ --> / NOM **EP** (ADJ*) / NOM **EP** (ADJ*) **AT** (GADJ)
GADJ --> / ADJ **PREP** INF / ADJ **PREP** CP
ARG --> / INF / NOM **EP** (ADJ*)
NOM --> / SUB / SUB **EP** SUB
ADJ --> / adjectif qualificatif / adjectif numéral / participe passé
SUB --> / substantif
INF --> / infinitif
PREP--> / **D** / **ND**
EP --> / /
AT --> / /
D --> / ***de*** / ***de*_det**
ND --> / **P** / **P_det**
P --> / toute autre préposition que ***de***

Les relations structurant les GC sont représentées par des non-terminaux en gras. Seules les relations prépositionnelles seront conservées explicitement dans les GC. Les symboles **EP** et **AT** qui signifient respectivement EPithète (antérieur ou postérieur) et ATtribut, précisent la nature de la relation de qualification adjectivale prise en compte lors de la reconnaissance du groupe conceptuel, relations qui ne seront pas explicitées dans les GC.

Il convient de préciser les deux points suivants :

-1) le groupe adjectival, GADJ, n'est reconnu qu'à partir d'un adjectif en position d'attribut. On se limite à cette seule possibilité de reconnaissance du groupe adjectival, afin de ne pas être confronté à la complexité des ambiguïtés de structure liées au problème du point de rattachement du complément prépositionnel : soit sur le nom qualifié, soit sur l'adjectif qualifiant.

Exemple :

"Le programme entier est difficile pour les étudiants. "

Dans ce cas, la position d'attribut de l'adjectif " *difficile* ", permet de déterminer sans ambiguïté que le complément prépositionnel se rattache à cet adjectif, pour former un groupe adjectival :

GADJ --> (*difficile* ***pour_*det** *étudiant*)

Ce qui nous donne le Groupe Conceptuel suivant :

GC --> *programme* (*entier*) (*difficile* ***pour_*det** *étudiant*)

NOM ADJ GADJ

-2) la prise en compte, lors de la reconnaissance de groupes conceptuels structurés par des relations de qualifications prépositionnelles enchaînées, de la seule relation de qualification adjectivale épithète antérieur (sauf pour le premier nom pour lequel la relation de qualification adjectivale épithète postérieur est également prise en compte), permet d'une part, de déterminer sans ambiguïté le nom qualifié pour chaque adjectif retenu, et d'autre part supprime la possibilité des groupes adjectivaux épithètes.

Exemple :

dans la phrase suivante, la détermination du nom qualifié par l'adjectif "*comptable* " en position d'épithète postérieur ne peut se faire automatiquement sans critères sémantiques :

" *La solution du problème de gestion comptable pour les premières années était évidente* .",

de même on ne peux pas décider si l'on est en présence d'un groupe adjectival avec :

"*comptable pour les premières années* ".

On peut rencontrer également le même type de phrases avec les déterminations inverses :

" *La solution du problème de gestion facile pour les premières années était délicate pour les secondes années.* ".

En conséquence, nous avons décidé d'ignorer les adjectifs en position d'épithète postérieur s'ils sont suivis d'un complément prépositionnel.

Par contre, dans les deux cas on est en présence de groupes adjectivaux en position d'attribut :

(*évident*) et (*délicat* ***pour_*det** *année* (*second*))

que l'on peut rattacher sans ambiguïté au premier nom du syntagme nominal.

Dans la première phrase on reconnaîtra le groupe conceptuel suivant :

(*solution* (*évident*) ***de_*det** (*problème* ***de*** *gestion*))
***pour_*det** *année* (*premier*)

Dans la deuxième on reconnaîtra :

(*solution* (*délicat* ***pour_*det** *année* (*second*)
de_*det** (*problème* ***de *gestion*))
***pour_*det** *année* (*premier*)

2.3. Classification des Groupes Conceptuels

La syntaxe-cible que nous venons de définir, permet de distinguer trois catégories de GC, en fonction de leur longueur et de leur complexité de structure.

Cette distinction sera nécessaire pour la sélection, lors de la phase d'indexation, des groupements jugés représentatifs d'un texte [KERK 84]. Pour réaliser cette sélection, la stratégie employée consiste à considérer d'abord les groupes conceptuels les plus longs. S'ils ne sont pas retenus, une cassure syntaxique sera alors effectuée, suivant cette distinction, pour considérer les sous groupes issus du groupement rejeté.

Ces trois catégories de GC sont constituées par les " **mots isolés** " ou GCM, les " **Groupes Conceptuels Simples** " ou GCS, et les

" **Groupes Conceptuels Complexes** " ou GCC, que nous présentons ci-après :

-1) **Le mot isolé** ou GCM, dont la syntaxe est :

GCM --> / NOM / INF

Le mot isolé peut être en fait soit un mot simple, SUB ou INF, soit un nom composé, ce peut être le cas d'un SUB s'il est connu du système (contenu dans le dictionnaire), et c'est le cas pour la structure SUB **EP** SUB.

On considère la structure SUB **EP** SUB comme un mot isolé (nom composé). Les deux substantifs évoquent dans ce cas une image unique et non la réunion de deux images distinctes. La langue française est particulièrement riche en noms composés de ce type :

" *Plan Rocard* " " *bleu ciel* " " *télévision couleur* "

Le deuxième substantif est alors souvent équivalent à un adjectif, dans le sens où il qualifie le premier. Néanmoins, cette juxtaposition est plus unificatrice que dans le cas de l'adjectif épithète. Les grammairiens expliquent la formation de ces noms composés par l'élision d'une préposition, ce qui prouverait la "naissance" de nouveaux concepts. Les exemples donnés ci-dessus provenant alors des expressions avec préposition suivantes :

" *Plan de Rocard* " " *bleu du ciel* " " *télévision en couleur* "

La structure SUB **EP** ADJ, peut être considérée comme un mot isolé dans de nombreux cas en français :

" *carte bleue* " " *feu rouge* " " *belle époque* "

Ces cas ne pouvant pas être décelés sur des critères strictement syntaxiques, aussi avons nous pris le parti de ne pas les considérer automatiquement comme des noms composés.

-2) **Le groupe conceptuel simple** ou GCS, dont la syntaxe est :

GCS --> / NOM **EP** ADJ / GCM **PREP** GCM

On peut remarquer que certaines structures permises par la syntaxe-cible ne seront jamais rencontrées en français telles que :

NOM **D** INF (avec **D** --> *du* , *des*)

cela n'est pas gênant, dans la mesure où notre but est ici de définir une représentation interne de structures présentes dans les textes.

Voici quelques exemples de groupes conceptuels simples :

carte (*postal*) , *maison* ***à*** *vendre* , *école* ***de*** *architecture*

-3) **Le groupe conceptuel complexe** ou GCC

Il s'agit de tout autre groupe conceptuel (GC), n'appartenant pas aux deux catégories précédentes.

Voici quelques exemples de groupes conceptuels complexes :

" *l'école de filles de la commune* "
---> (*école* ***de*** *fille*) ***de***_**det** *commune*

" *le programme est difficile pour les élèves* "
---> *programme* (*difficile* ***pour***_**det** *élève*)

" *la belle petite chienne du boucher* "
---> *chienne* (*beau petit*) ***de***_**det** *boucher*

La définition des groupes conceptuels reste extensible. En particulier lors d'un développement ultérieur, l'ajout d'un sous-ensemble des syntagmes verbaux reste possible. Cet ajout deviendrait nécessaire si l'on s'orientait vers des approches fondées sur une représentation plus élaborée de la connaissance (actuellement les liaisons sémantiques, par exemple, ne sont pas explicitées; elles restent implicites dans la forme du Groupe Conceptuel).

2.4. Connexions entre les Groupes Conceptuels

Une indexation fondée uniquement sur les groupes conceptuels perd une partie du sens, ou de la connaissance, contenus dans la phrase. C'est pourquoi, lors du processus d'indexation, les groupes conceptuels sélectionnés pourront être connectés par un ensemble de relations reconnues au cours de l'analyse textuelle. Ces relations donnent une interprétation partielle des syntagmes verbaux.

Dans un premier temps, toujours pour effectuer des reconnaissances aisées en évitant la complexité des ambiguïtés des structures syntaxiques, nous nous sommes limités à trois types de relations :

- Les relations "prépositionnelles"

- Les relations "auxiliaires"

- La relation "de proximité"

2.4.1. Les relations prépositionnelles

Ces relations sont introduites par les prépositions suivant directement les verbes conjugués. Elles traduisent la notion de complément d'objet indirect.

Les relations prépositionnelles permettent donc de représenter le lien syntaxique existant entre le groupe conceptuel extrait du groupe nominal sujet et le groupe conceptuel extrait du groupe nominal complément d'objet indirect.

Certaines prépositions sont porteuses d'informations sémantiques intéressantes (cf. II.2.4.4), c'est pourquoi nous différencions les relations en fonction de la préposition.

Exemples :

" *L'oiseau niche sur la branche.* "

---> *oiseau* **RP_*sur*** *branche*

" *La bûche brûle dans la cheminée* "
---> *bûche* **RP_*dans*** *cheminée*

2.4.2. Les relations auxiliaires

Ces relations sont introduites par les auxiliaires " *avoir* " et " *être* " lorsqu'ils ne sont pas suivis d'une préposition, c'est-à-dire lorsqu'ils n'introduisent pas de complément d'objet indirect (cas précédent). Elles traduisent généralement des relations d'appartenance ou de possession (auxiliaire " *avoir* ") et de généricité (auxiliaire " *être* ").

Exemples :

" *Un lapin a deux grandes oreilles.* "
---> *lapin* **AUXA** *oreille* (*deux grand*)

" *La rose est une fleur.* "
---> *rose* **AUXE** *fleur*

Ce type de relations peut être très intéressant pour la structuration automatique ou semi-automatique d'une base de connaissances. La relation **AUXE** peut aider à établir une relation sémantique de type hiérarchique, telle que la généricité.

2.4.3. La relation de proximité

Cette relation permet de connecter un groupe conceptuel issu d'un groupe nominal sujet avec un groupe conceptuel issu d'un groupe nominal complément d'objet direct.

Exemple :

" *L'enfant mange un gâteau.* "
---> *enfant* **PROX** *gâteau*

Il serait tout à fait envisageable, pour une application donnée, de définir un ensemble de verbes "intéressants" pour l'application, afin de diversifier cette relation **PROX**, comme il a été fait pour les relations prépositionnelles.

2.4.4. Intérêt des relations entre Groupes Conceptuels

Ces relations permettent de représenter un noyau d'informations extrait lors de l'analyse du texte, automatiquement et sans en interpréter le sens.

Le choix des relations prépositionnelles et de la relation générique doit permettre, au cours d'une phase ultérieure, par une étude de leurs propriétés et des rapports exprimés par les prépositions, de réaliser des inférences permettant de compléter le travail classique effectué au niveau du thésaurus.

Les prépositions peuvent exprimer, pour relier et subordonner deux éléments d'une phrase, de nombreux rapports [GREV 80], tels que :

- le lieu, (" *dans, chez, sous, devant, etc...*)
- le temps, (" *depuis, pendant, durant, etc...*)
- la cause, (" *attendu, pour, à cause de, etc...*)
- la manière, (" *avec, selon, sans, etc...*)
- etc...

Nous ne traitons pas pour l'instant le cas de la négation, mais nous n'extrayons pas de relation lorsque l'on rencontre une particule négative. Il est à noter que le traitement des formes négatives requiert une attention particulière, car dans un certain nombre de cas, elles ne peuvent pas se traduire simplement par la négation de la relation extraite (qui peut d'ailleurs ne rien signifier).

2.5. Extensions possibles

Une des principales extensions envisageables au niveau de la définition de la syntaxe des GC est la prise en compte totale des syntagmes verbaux.

Nous venons de voir que grâce aux relations (prépositionnelles, auxiliaires, de proximité), déjà définies entre les Groupes Conceptuels, nous avons en quelque sorte réalisé une première interprétation de ces syntagmes verbaux. Cette approximation se justifie pour un traitement linguistique préparant une phase d'indexation, en extrayant les entités conceptuelles contenues dans les syntagmes nominaux du texte traité.

Les premiers résultats obtenus en testant nos méthodes sur une syntaxe plus réduite nous ont conforté dans notre démarche. Ces résultats ont servi à la réalisation du processus d'indexation automatique développé par D. Kerkouba [KERK 84].

Il en va tout autrement pour la phase d'interrogation qui nécessite un traitement linguistique, non plus du texte mais de la requête, plus poussé. En effet, lors de l'interrogation les formes verbales peuvent avoir un degré de signification (contenu sémantique) essentiel. Une étude dans ce sens est menée au sein de l'équipe, et a abouti à la réalisation d'une première maquette [NIE 85], prolongée par une application dans le cadre du projet RIME [NIE 90].

Une autre extension étudiée est la résolution des références pronominales, et plus particulièrement le traitement des pronoms relatifs, souvent utilisés en français ne serait-ce que pour des questions de stylistique. Cette résolution semble contradictoire avec nos impératifs d'efficacité, fixés par l'évolution du domaine vers la constitution de bases de données textuelles, où le texte est entièrement saisi. D'où notre proposition d'outils spécifiques (traitements partiels) pour des corpus importants. Un compromis intéressant est envisageable en se limitant à la résolution de certaines références peu coûteuses, en particulier lors de l'emploi de pronoms relatifs dans des expressions syntaxiquement figées.

Enfin certaines tournures stylistiques simples pourraient également faire l'objet de traitements spécifiques :

Exemple :

Dans la phrase suivante, nous retrouvons un groupe adjectival non continu :

" *Pour les premières années, la solution du problème de gestion comptable était évidente .* "

que l'on pourrait reconstituer :

" *évidente pour les premières années* "

3. Le processus d'extraction des groupes conceptuels

Le processus d'extraction des groupes conceptuels, basé sur la syntaxe cible que nous venons de définir, est réalisé par un transducteur d'états finis. L'objectif de ce processus est de produire à faible coût, à partir du résultat d'une analyse de surface de la langue naturelle, un ensemble de groupes conceptuels. Ces GC seront soumis ensuite au processus d'indexation automatique proprement dit (sélection des GC les plus représentatifs du texte traité avec éventuellement cassure syntaxique) pour produire les termes d'indexation.

L'entrée de ce processus d'extraction est constituée par un réseau de solutions morphologiques partiellement désambiguïsé produit par l'analyseur détaillé au cinquième chapitre. Une présentation formelle de ce réseau est donnée au quatrième chapitre (cf. IV.3.6.1).

La transduction ne s'effectue que sur les portions désambiguïsées du réseau (portions linéaires). Les groupes conceptuels étant construits à partir des syntagmes nominaux, un des impératifs conditionnant la qualité de cette transduction, est de disposer en entrée d'un réseau de solutions morphologiques où les syntagmes nominaux sont fortement désambiguïsés.

Enfin, ce transducteur doit pouvoir évoluer facilement si l'on veut conserver le caractère extensible de la syntaxe cible définissant les groupes conceptuels.

Cette construction s'est effectuée en plusieurs étapes :

Tout d'abord, une syntaxe cible simplifiée, limitée aux GCM (mots isolés) et aux GCS (groupes conceptuels simples) que nous avons défini précédemment (cf. 2.3.) a donnée lieu a un premier transducteur.

Réalisé en PROLOG, il a servi à l'expérimentation du processus d'indexation dynamique réalisé par D. Kerkouba ; on peut trouver les résultats de cette expérimentation dans [KER 85].

Le vocabulaire d'entrée est constitué par l'ensemble des catégories grammaticales défini dans le modèle linguistique. Le vocabulaire de sortie est

constitué des catégories des éléments participants à la définition des GC et des symboles représentant les relations retenues.

Ce premier transducteur a été complété ensuite de manière à prendre en compte l'ensemble de la syntaxe cible des GC, à l'exception des groupes adjectivaux. Pour cela, il a suffit de définir les transitions complémentaires à partir de l'ensemble des états finaux du premier transducteur.

La table définissant la fonction de transition est donnée ci-après. Certaines catégories du vocabulaire d'entrée ont été regroupées, pour plus de lisibilité, dans des classes plus générales (tout comme dans la syntaxe cible) :

S	pour	SUB	= { SUBC, SUBP }
A	pour	ADJ	= { ADJQ, ADJO, ADJC, VBPA }
D	pour	D	= { *de* }
P	pour	P	= { PREP }
T	pour	det	= {ARTD, ARTI, ADJP, ADJD }
I	pour	INF	= { VBIF }
≠	pour	les autres catégories	

La signification de ces catégories est donnée en annexe (cf. annexe 1).

Chaque case de cette table est constituée d'un triplet de la forme :

$$< n , i , a >$$

où **n** est le prochain état, **i** une inscription à effectuer sur la bande de sortie, et **a** un ensemble d'actions à exécuter soit en entrée soit en sortie.

L'état initial est l'état "0", l'état final provoque après l'exécution des actions réalisées, l'écriture de la sortie.

ETAT	S	A	D	P	T	I	≠
0	36, S,	42, ..A,	0, ,	0, ,	0, ,	1, I,	0, ,
1	F, ,	F, ,	2, D,	3, P,	F, ,	F, , =	F, ,
2	4, S,	5, ..A,	F, , p	F, , p	6, T,	F, I,	F, ,p
3	7, S,	8, ..A,	F, , p	F, , p	9, T,	F, I,	F, , p
4	10, S,	11, A,	2, D,	12, P,	F, ,	F, , =	F, ,
5	13, -S,	5, A,	F, , p	F, , p	F, , p	F, , p=	F, , p
6	4, S,	5, ..A,	F, , p	F, , p	F, , p	F, , p=	F, , p
7	14, S,	15, A,	16, D,	F, ,	F, ,	F, ,=	F, ,
8	17, S,	8, A,	F, , p	F, , p	F, , p	F, , p=	F, , p
9	7, S,	8, ..A,	F, , p	F, , p	F, , p	F, , p=	F, , p
10	F, , =	11, A,	2, D,	12, P,	F, ,	F, , =	F, ,
11	F, , =	11, A,	F, ,	F, ,	F, ,	F, ,=	F, ,
12	18, S,	19, ..A,	F, , p	F, , p	20, T,	F, I,	F, , p
13	21, -S,	22, A,	2, D,	12, P,	F, ,	F, , =	F, ,
14	F, , =	23, A,	16, D,	24, P,	F, ,	F, , =	F, ,
15	F, , =	15, A,	F, ,	F, ,	F, ,	F, , =	F, ,
16	25, S,	26, ..A,	F, , p	F, , p	27, T,	F, I,	F, , p
17	23, -S,	23, A,	16, D,	24, P,	F, ,	F, , =	F, ,
18	28, S,	F, ,	F, ,	F, ,	F, ,	F, , =	F, ,
19	29, S,	19, A,	F, , p	F, , p	F, , p	F, , p=	F, , p
20	18, S,	19, ..A,	F, , p	F, , p	F, , p	F, , p=	F, , p
21	F, , =	22, A,	2, D,	12, P,	F, ,	F, , =	F, ,
22	F, , =	22, A,	F, ,	F, ,	F, ,	F, , =	F, ,
23	F, , =	23, A,	16, D,	24, P,	F, ,	F, , =	F, ,
24	30, S,	31, ..A,	F, , p	F, , p	32, T,	F, I,	F, , p
25	33, S,	F, ,	16, D,	F, ,	F, ,	F, , =	F, ,
26	34, -S,	26, A,	F, , p	F, , p	F, , p	F, , p=	F, , p
27	25, S,	26, ..A,	F, , p	F, , p	F, , p	F, , p=	F, , p
28	F, , =	F, ,	F, ,	F, ,	F, ,	F, , =	F, ,
29	F, -S,	F, ,	F, ,	F, ,	F, ,	F, , =	F, ,
30	F, S,	F, ,	F, ,	F, ,	F, ,	F, , =	F, ,
31	30, -S,	31, A,	F, , p	F, , p	F, , p	F, , p=	F, , p
32	30, S,	31, A,	F, , p	F, , p	F, , p	F, , p=	F, , p
33	F, , =	F, ,	16, D,	F, ,	F, ,	F, , =	F, ,
34	35, -S,	F, ,	16, D,	F, ,	F, ,	F, , =	F, ,
35	F, , =	F, ,	16, D,	F, ,	F, ,	F, , =	F, ,
36	37, S,	36, A,	39, D,	40, P,	F, ,	F, , =	F, ,
37	F, , =	37, A,	39, D,	40, P,	F, ,	F, , =	F, ,
38	37, -S,	37, A,	39, D,	40, P	F, ,	F, , =	F, ,
39	4, S	5, ..A,	F, , p	F, , p	6, T,	F, I,	F, , p
40	41, S,	8, ..A,	F, , p	F, , p	9, T,	F, I,	F, , p
41	F, S,	F, ,	F, ,	F, ,	F, ,	F, , =	F, ,
42	38, -S,	42, A,	F, , a	F, , a	F, , a	F, , a=	F, , a

- le prochain état est désigné par son numéro, ou par la lettre F indiquant l'état final.

- l'inscription à effectuer en sortie peut être :
  - soit vide,
  - soit la catégorie lue en entrée

  dans le cas des substantifs, si le code est précédé d'un "-", cela signifie que ce code doit être inscrit à l'emplacement repéré par un ".",

  dans le cas des adjectifs, si le code est précédé de "..", cela signifie que l'on réserve une ou deux places pour un substantif qui doit suivre.

- les actions à exécuter sont de quatre types :
  - un symbole "=" signifie que l'on n'avance pas en lecture;
  - un symbole "p" signifie que l'on dépile tous les symboles en sortie jusqu'à un "P" ou un "D" inclus;
  - un symbole "a" signifie que l'on efface la sortie;
  - sinon, on progresse d'une catégorie en lecture.

C'est cette version du transducteur qui a été réalisée et expérimentée. Les résultats, présentés au chapitre VI permettent de valider cette démarche.

La reconnaissance des groupes adjectivaux et des relations entre groupes conceptuels nécessite la définition d'automates spécifiques, dont les états initiaux correspondent à l'état final "F", et dont les états finaux correspondent à l'état initial "0".

4. Conclusion

La définition des Groupes Conceptuels que nous venons de présenter représente un compromis entre le degré de structuration des concepts extraits des textes analysés, et la complexité de l'analyse linguistique mise en oeuvre à cet effet. Le choix des syntagmes nominaux n'est pas à cet égard original, mais les relations proposées entre les G.C., permettent sans augmenter la complexité

de l'analyse, d'envisager à partir notamment des extensions concernant les syntagmes verbaux (cf. III.2.5), de s'orienter vers une une aide à la structuration de connaissances (construction assistée de thésaurus).

L'extraction des Groupes Conceptuels repose sur l'exploitation par un transducteur d'états finis, d'un réseau de solutions morphologiques fortement linéarisé. La linéarisation complète n'étant pas nécessaire, l'objectif de l'analyseur de surface développé, est de concentré cette linéarisation sur les portions de réseau susceptibles de contenir des Groupes Conceptuels potentiels, donc sur les portions de réseau susceptibles de contenir des groupes nominaux.

Le modèle linguistique que nous présentons dans le chapitre suivant, tout en restant proche des modèles classiques, comporte dans cette optique, certaines particularités notamment en ce qui concerne le choix des classes syntaxiques.

CHAPITRE IV

PLAN DU CHAPITRE IV

MODELE LINGUISTIQUE ET FORMALISME UTILISE

MODELE LINGUISTIQUE ET FORMALISME UTILISES

1. Introduction

Nous présentons dans ce chapitre la spécification de l'analyseur qui nous permet de transformer un texte écrit en langue naturelle, en une forme exploitable par le processus d'extraction des groupes conceptuels, que nous avons présentés au chapitre précédent (cf. III.3). La stratégie mise en oeuvre est une analyse de surface intégrant des composants d'essences morphologique et syntaxique. C'est cette coopération morphologie-syntaxe qui nous permet en particulier de traiter les mots inconnus sans interrompre le processus d'analyse (utilisation d'un module de catégorisation automatique). Cette présentation va s'effectuer en deux parties :

-1) Présentation du modèle linguistique :

nous commencerons par définir le modèle linguistique sur lequel repose cette analyse. Nous préciserons à cette occasion l'ensemble des données linguistiques initiales fournies au système. La détermination de cet ensemble initial est un des aspects importants de notre travail, qui

conditionne l'aspect acquisition automatique d'informations linguistiques, à la base des processus de catégorisation automatique et d'enrichissement du vocabulaire. En effet, cet ensemble de données linguistiques initiales contient entre autres l'ensemble des irrégularités grammaticales du français (recensées le plus exhaustivement possible), ce qui nous autorise à inférer un comportement grammatical "régulier" lors de la rencontre d'un mot inconnu en cours d'analyse.

-2) Présentation formelle :

nous présenterons ensuite le formalisme permettant d'introduire les données manipulées par les différentes composantes de l'analyseur. Nous nous efforcerons dans cette présentation formelle d'interpréter les résultats obtenus aux différentes étapes du processus. C'est cette étude formelle qui nous permettra de justifier l'adéquation de l'analyse de surface mise en oeuvre, aux objectifs exposés au chapitre précédent (cf. III.2)

Nous conclurons enfin, en exposant la problématique de l'analyseur que nous développerons au chapitre suivant.

2. Définition d'un modèle linguistique

Avant toute analyse d'une langue naturelle, il est nécessaire de définir le modèle linguistique de référence (par rapport auquel se situe l'analyse).

Par modèle linguistique, nous entendons une classification des mots de la langue, en fonction de leur rôle syntaxique, ou des différentes positions que pourront prendre ces mots dans l'ordonnancement d'une phrase de la langue. Ces classes se nomment traditionnellement classes syntaxiques élémentaires, ou encore catégories grammaticales. Pour chaque catégorie grammaticale, un ensemble de variables, qualifiées de grammaticales, pourront permettre de préciser un ensemble d'attributs grammaticaux (ou encore de valeurs grammaticales), qui constituent les renseignements linguistiques associés aux éléments de ces classes :

Exemple :

Un NOM (catégorie des substantifs) pourra avoir deux variables grammaticales associées : le "genre", notée GNR, et le "nombre" notée NBR, qui pourront, respectivement, prendre leurs valeurs parmi les ensembles suivants :

{ masculin , féminin } et { singulier , pluriel }

" ***hommes*** " ---> NOM, GNR = {masculin},
NBR = {pluriel}

" **enfant** " ---> NOM, GNR = { masculin , féminin },
NBR = { singulier }

Dans la littérature spécialisée, pour une même langue naturelle, il existe presque autant de modèles linguistiques que de linguistes : la définition d'un modèle linguistique est étroitement liée à l'utilisation que l'on veut en faire, au niveau d'analyse que l'on veut obtenir. Nous en avons vu quelques exemples au chapitre précédent, au travers des quelques systèmes présentés.

Nous n'échapperons pas à cette "règle", et le modèle que nous avons défini, est dépendant aussi bien de la stratégie d'analyse que nous mettons en oeuvre, que de l'application envisagée : l'aspect positionnel de la langue française a été déterminant dans la conception de la stratégie d'analyse employée, et a fortement influé sur les choix effectués pour la définition des catégories grammaticales utilisées. Néanmoins, cela ne constitue en aucun manière une limitation de notre système, car nous allons voir que les choix effectués peuvent permettre une transposition relativement simple des données linguistiques dans un autre modèle, car nous avons dans l'ensemble conservé la classification de base de la grammaire traditionnelle du français.

Pour définir notre modèle, nous avons essentiellement pris en compte deux aspects de nos motivations :

- la reconnaissance de structures particulières de la langue naturelle (en l'occurrence les syntagmes nominaux),

- le traitement des mots inconnus et la possibilité d'enrichissement automatique du vocabulaire.

Pour cela, nous avons considéré d'une part, les classes syntaxiques dont la liste exhaustive des éléments peut être établie a priori (ces classes sont dites

fermées), et dont le rôle syntaxique permet de faciliter l'identification de structures particulières de la langue naturelle (syntagmes nominaux, verbaux, etc...), et d'autre part les classes dites ouvertes, pour lesquelles il est impossible de dresser des listes exhaustives. Ce sont les classes ouvertes qui feront l'objet d'apprentissage de mots nouveaux.

Ainsi nous pourrons par cette diversification affiner le rôle joué par les "indicateurs de structures syntagmatiques" appartenant à ces classes fermées, sans compliquer le traitement des mots inconnus, dont une étape essentielle est l'attribution automatique de catégories grammaticales potentielles (qui ne pourront être que celles des classes ouvertes).

2.1. Notions de classe fermée et de classe ouverte

Nous appelons classe fermée, toute classe syntaxique du français dont l'énumération exhaustive des éléments est possible; par opposition aux classes dites "ouvertes", dont le nombre d'éléments est forcément fini (l'ensemble des mots de la langue française est fini), mais pour lesquelles l'énumération exhaustive est pratiquement impossible, du fait que ces classes sont continuellement enrichies de mots nouveaux (évolution de la langue, acquisition de mots d'origine étrangère, intégration de termes scientifiques et techniques).

Exemple :

Les articles définis forment une classe fermée dont la liste exhaustive des éléments est : " *le, la, l', les* ".

Par contre les substantifs, les adjectifs qualificatifs, les adverbes, les verbes forment des classes ouvertes.

Outre ces classes syntaxiques, nous pouvons déterminer pour les langues naturelles certains sous-ensembles fermés des classes ouvertes possédant des propriétés grammaticales singulières. Cela est particulièrement le cas en français. En effet, la langue française possède un grand nombre de règles de grammaire dont une des caractéristiques semble être les listes d'exceptions plus ou moins nombreuses qui y sont attachées : règles d'accord, de conjugaison, etc...

L'ensemble de ces irrégularités grammaticales provient d'un "héritage" de la langue. Ces irrégularités sont donc bien connues et nous pouvons par

conséquent les considérer également comme des classes fermées du français (sous ensembles facilement énumérables des classes ouvertes). C'est ainsi que par exemple, tous les verbes appartenant au troisième groupe de conjugaison sont énumérables et répertoriés :

" *Ainsi se trouve exactement circonscrite cette conjugaison morte qui par sa complexité et ses singularités constitue la difficulté majeure du système verbal français* . " [BESC 80].

On peut y adjoindre les auxiliaires " *être* " et " *avoir* ", ainsi que le verbe " *haïr* " qui bien qu'appartenant au deuxième groupe de conjugaison, comporte certaines particularités.

En ce qui concerne les substantifs et les adjectifs qualificatifs, nous avons répertorié les formations irrégulières du pluriel. Nous aurions pu également prendre en compte les féminins irréguliers, nous reviendrons sur ce choix lors de la présentation de l'enrichissement automatique du vocabulaire.

Pour les adverbes nous avons recensé uniquement ceux qui ne se terminent pas en "*ment* " que nous considérons comme des "irrégularités".

Nous avons donc recensé en résumé, pour ce qui concerne les classes fermées, les éléments suivants :

- les constituants des classes fermées traditionnelles : les déterminants (articles, adjectifs déterminatifs), les pronoms (personnels, possessifs, etc...), les prépositions et les conjonctions,

- et les principales irrégularités grammaticales recensées, constituées par : les verbes du troisième groupe de conjugaison, les adverbes ne se terminant pas en " *ment* ", les substantifs et adjectifs qualificatifs ayant un pluriel irrégulier.

En fait, ces classes fermées nous fournissent une substance pour l'initialisation du dictionnaire. Cette initialisation est nécessaire pour la stratégie d'enrichissement automatique du vocabulaire que nous mettons en oeuvre, et en particulier pour l'attribution automatique des catégories grammaticales potentielles et des valeurs grammaticales associées (cf. V.6.1).

Les classes ouvertes comportent les substantifs, les adjectifs qualificatifs, les adverbes et les verbes dont le comportement est régulier.

La tendance actuelle en français lors de l'acquisition ou de la création de mots nouveaux, qui proviennent pour la plupart des divers domaines techniques ou des langues étrangères, est une régularisation grammaticale systématique. En conséquence on peut penser que l'évolution ultérieure de la langue française restera parfaitement compatible avec notre démarche, en particulier en ce qui concerne notre stratégie d'enrichissement automatique du vocabulaire.

Ce recensement des irrégularités grammaticales et orthographiques, s'il n'est pas exhaustif (il ne prend pas en compte les féminins irréguliers des substantifs et adjectifs, ou les orthographes multiples de certains mots), permet néanmoins dans une première approche un fonctionnement tout à fait acceptable de l'analyseur dans le contexte de l'application projetée.

A cet égard l'intégration des résultats de cette nature fournis par la communauté linguistique constitue une possibilité intéressante de développement de cette étude. On peut citer par exemple le groupe de linguistes du Laboratoire d'Automatique Documentaire et Linguistique (LADL), dirigé par Maurice GROSS, qui s'applique au travers d'une étude systématique de la syntaxe du français, à établir des listes exhaustives d'exceptions aux règles habituelles de grammaire du français (pluriels irréguliers, féminins irréguliers, etc...).

On peut regretter toutefois, que ce travail important pour toute application visant à une analyse automatique de la langue naturelle de portée générale (non limitée à un sous langage propre à une application dans un domaine particulier), n'ait pas été entrepris beaucoup plus tôt par la communauté des linguistes. Il est vrai que ce type d'investigation demande à la fois de fortes compétences linguistiques et un grand potentiel humain.

Nous avons pu vérifier à nos dépens, la complexité d'une telle entreprise. Il est bon néanmoins de considérer que la constitution de ces listes exhaustives est déjà en partie réalisée au LADL, ce qui permet de conforter notre démarche et d'en confirmer la faisabilité.

2.2. Choix des catégories grammaticales

Parmi les renseignements grammaticaux associés aux formes contenues dans le dictionnaire figurent les catégories grammaticales. Ces catégories nous permettent d'identifier grammaticalement les formes rencontrées.

Nous avons retenu à peu près les catégories classiques de la grammaire française, en effectuant quelques choix que nous allons justifier.

Tout d'abord, il est à remarquer qu'un certain nombre de mots appartenant à des catégories grammaticales différentes jouent à peu près le même rôle syntaxique. Nous pouvons citer par exemple les déterminants. Lorsqu'en plus ces mots constituent des homographies (même forme, mais catégories grammaticales différentes), il est tentant de les regrouper sous une même catégorie grammaticale, surtout dans le cas d'une analyse syntaxique partielle où la différenciation des rôles syntaxiques est moins fine :

Exemple :

adjectifs interrogatifs et adjectifs exclamatifs.

Néanmoins pour les classes fermées, nous avons diversifié les catégories grammaticales lorsque cela présente un intérêt au niveau de la syntaxe et plus particulièrement pour la levée des ambiguïtés grammaticales. Cette diversification ne portant que sur les classes fermées ne complique aucunement le processus d'enrichissement automatique du vocabulaire, car ces classes sont définies une fois pour toutes lors de l'initialisation du dictionnaire. L'intérêt de cette diversification est de permettre un affinage des relations positionnelles par la distinction des figures d'ambiguïté qu'elle provoque. L'exemple suivant en est une parfaite illustration :

Exemple :

La diversification des catégories de déterminants, pour entre autres les catégories des articles définis, des adjectifs possessifs, des adjectifs démonstratifs, des articles indéfinis, etc...

Pour les éléments des classes ouvertes, nous avons déterminé neuf catégories grammaticales :

SUBC substantifs communs
SUBP substantifs propres
ADJQ adjectifs qualificatifs
ADVB adverbes
VBIF verbes à l'infinitif
VBPA verbes au participe passé
VBPR verbes au participe présent
VBCJ verbes conjugués
ABRV abréviations et sigles

Nous avons diversifié la classe syntaxique des verbes, afin de pouvoir distinguer les rôles syntaxiques différents joués par ces sous-ensembles et notamment les relations positionnelles qu'ils entretiennent avec les autres catégories grammaticales. Nous avons vu au chapitre précédent (cf. III.2.1), que les verbes à l'infinitif représentent une forme nominale du verbe, et que les verbes au participe présent ou passé, ont une fonction adjectivale. De plus ces trois catégories grammaticales ont un ensemble de terminaisons morphologiques qui permettent de les distinguer facilement des verbes conjugués (sauf pour certains participes passés).

De même la distinction entre les substantifs communs et les substantifs propres sera aisée puisque se faisant simplement sur le caractère majuscule ou minuscule de la première lettre de la forme analysée.

Enfin nous avons une catégorie abréviations et sigles, qui représente également une classe ouverte. La reconnaissance d'un sigle ou d'une abréviation se fera dans des cas simples (forme inconnue tout en majuscule, alternance de lettres majuscules et de points, forme inconnue suivie d'un point sans que la forme suivante ne commence par une majuscule).

Nous avons retenu ces deux dernières catégories, SUBP et ABRV, dont le comportement syntaxique des éléments est similaire de ceux de la catégories des substantifs communs, du fait de la relative aisance de leur reconnaissance à partir de critères simples (cf. V.3.2.3). Cette reconnaissance peut être particulièrement intéressante dans certaines applications et n'entraîne pas de complexité supplémentaire significative du processus d'analyse.

Nous avons également été tentés d'effectuer certains regroupements de catégories grammaticales après coup, c'est-à-dire en cours d'expérimentation, lorsque la diversification nous est apparue inutile pour le niveau de la phase d'analyse textuelle recherchée. C'est ainsi que nous aurions pu supprimer la catégorie grammaticale notée ARTC, comprenant les articles contractés : " *au, aux, du, des* ", pour la fondre avec la catégorie PREP regroupant les prépositions. Mais nous perdions ainsi l'information concernant la présence d'un article défini (indicateur de groupe nominal). Or outre le rôle d'indicateur de structure important, cette présence d'un article défini intervient également comme un élément déterminant dans le processus d'indexation, pour effectuer certaines cassures syntaxiques [KERK 84]. On peut rappeler que ces articles contractés proviennent de la combinaison des propositions " *à* " et " *de* ", avec les formes " *le* " et " *les* " de l'article défini.

Par contre le rôle syntaxique particulier joué par la préposition " *dans* " qui introduit toujours un groupe nominal (et jamais un groupe verbal), se

rapprochant ainsi des articles contractés, nous a permis de "créer" de façon artificielle les nouveaux déterminants composés :

" *dans le* ", " *dans la* ", " *dans les* ", " *dans l'* "

que nous rangeons avec les articles contractés dans la nouvelle catégorie grammaticale des articles composés, que nous codons ARTC. Ainsi par cet artifice, après la préposition " *dans* ", nous évitons l'ambiguïté classique entre un article défini et un pronom préverbal après une préposition :

Exemple :

l'expression " *pour le savoir* " a deux interprétations possibles :

préposition - pronom préverbal - verbe à l'infinitif

et

préposition - article défini - substantif commun

par contre l'expression " *dans le savoir* " n'en aura qu'une :

article composé - substantif

On pourra trouver en annexe (cf. annexe 1), la liste des catégories grammaticales que nous avons retenues dans un premier temps, car il est certain qu'en allant au bout du raisonnement que nous venons de mener, un certain nombre de modifications visant à préciser le rôle syntaxique de certains éléments particuliers, pourrait y être apportées, notamment en exploitant de manière systématique les renseignements que l'on peut trouver dans les travaux de linguistes concernant certaines structures particulières de la langue naturelle : on peut citer en particulier les travaux de M. Tutescu sur le groupe nominal en français, qui contiennent une mine de renseignements sur le rôle particulier des différentes classes de déterminants, exploitables de cette manière [TUTE 72].

2.3. Les variables grammaticales

Les variables grammaticales que nous utilisons, sont les variables traditionnelles pour le français. Elles sont au nombre de cinq :

- la variable grammaticale "genre" notée GNR, dont l'ensemble des valeurs possibles est :

{ masculin, féminin, { masculin, féminin } }

- la variable grammaticale "nombre" notée NBR, dont l'ensemble des valeurs possibles est :

{ singulier, pluriel, { singulier, pluriel } }

- la variable grammaticale "mode" notée MOD, dont l'ensemble des valeurs possibles est :

{ indicatif, subjonctif, conditionnel, impératif }

Les modes infinitif et participe sont représentés par des catégories grammaticales particulières.

- la variable grammaticale "temps" notée TMP, dont l'ensemble des valeurs possibles est :

{ présent, imparfait, passé simple, futur }

Les temps composés ne sont pas analysés en tant que tels, mais comme la succession d'un auxiliaire conjugué et d'un participe passé.

- la variable grammaticale "personne" notée PRS, dont l'ensemble des valeurs possibles est :

{ 1ère , 2ème , { 1ère , 2ème } , 3ème , { 1ère , 3ème} }

Cette variable est toujours utilisée avec la variable "nombre".

Les valeurs "doubles" de certaines variables grammaticales seront utiles dans la détermination des solutions morphologiques englobantes (cf.IV.3.2.1), ainsi que pour les vérifications d'accord grammatical déclenchées pendant le processus de filtrage syntaxique entre certaines formes (cf V.4.2).

On pourra trouver en annexe (cf. annexe 2), à quelles catégories grammaticales peuvent être associées chacune de ces variables. C'est l'ensemble des associations catégorie grammaticale - variables grammaticales qui forme la trame du modèle linguistique. C'est à partir de ces associations que l'on pourra définir un certain nombre de contraintes grammaticales, dont les principales sont les accords en genre, nombre, et personne.

Il est important de préciser qu'une catégorie grammaticale peut avoir plusieurs mêmes ensembles de variables grammaticales (avec des valeurs différentes) associés pour une forme, et qui correspondent à des interprétations grammaticales différentes de cette forme :

Exemple :

" *rédigions* " VBCJ (indicatif, imparfait, 1ère, pluriel)
(subjonctif, présent, 1ère, pluriel)

Dans le cas du mot " *rédigions* ", il ne s'agit que des deux interprétations possibles de deux formes conjuguées homographes du verbe " *rédiger* ", ne se différenciant que par les valeurs des variables grammaticales "mode" et "temps". Cette homographie est une constante de la conjugaison des verbes en français, pour ces valeurs grammaticales, et cette ambiguïté ne constitue pas un obstacle pour la reconnaissance du concept qui est le même dans les deux cas.

Par contre dans l'exemple suivant, la différence de genre pour le substantif " *page* ", permet de distinguer deux concepts totalement différents. On appelle classiquement une telle forme, une forme polysème.

Exemple :

" *page* " SUBC (masculin, singulier)
(féminin, singulier)

La reconnaissance de polysèmes sur des critères grammaticaux n'est pas toujours possible en français, il est nécessaire alors de considérer des critères contextuels de type syntaxique ou sémantique.

Exemple :

le substantif " *carte* " peut prendre plusieurs sens en français :

carte (*à jouer*),
carte (*postale*),
carte (*électronique*),
carte (*topographique*), etc...

Pour tous ces concepts différents, une analyse morphologique donnera pour le substantif " *carte* " la même interprétation dans le modèle linguistique :

SUBC (féminin, singulier).

Un des intérêts des groupes conceptuels sera de permettre cette distinction, en conservant une partie significative et discriminante du contexte des substantifs polysèmes, sans utilisation de critères syntaxiques ou sémantiques a priori. Cette distinction est particulièrement intéressante dans le contexte d'un Système de Recherche d'Informations puisqu'elle apporte une précision supplémentaire concernant le sens des concepts. Ce qui ne peut qu'entraîner une diminution du bruit relatif à une requête concernant un tel concept.

3. Spécification de l'analyseur de surface : formalisme utilisé

Nous donnons tout d'abord quelques définitions et propriétés formelles destinées à introduire et à expliciter le vocabulaire qui sera ensuite utilisé dans cette présentation.

Nous commençons par la définition des objets de base de ce formalisme, avant d'examiner les objets manipulés par chacun des trois composants essentiels de notre stratégie d'analyse, qui sont la morphologie, le filtrage syntaxique et la résolution des ambiguïtés grammaticales. Ces objets correspondent aux différentes structures de données construites à partir du texte initial, sur lesquelles travaillent ces modules.

3.1. Définition des objets de base

Nous présentons dans cette section un certain nombre de notions concernant les objets de base, permettant de cerner les éléments intervenant dans l'expression du contenu d'un texte.

-a) Lettre

Une lettre est un caractère appartenant à l'alphabet classique du français en minuscule ou en majuscule, ou la réunion d'un signe orthographique et d'un caractère pouvant former une voyelle

accentuée, ou le " c cédille". Si l'on désigne par LET, l'ensemble des lettres, on a :

LET = { a,b,c,...,z,A,B,C,...,Z,é,è,à,ù,â,ê,î,ô,û,ë,ï,ü,ç}

Remarque 1 :

Nous ne prenons pas en compte dans cette définition les signes orthographiques "œ" et "æ" que nous considérons comme les successions des caractères "o" et "e" d'une part, et "a" et "e" d'autre part.

Exemple :

La chaîne de caractères " *œil* " sera représentée par la chaîne "*oeil* ".

Remarque 2 :

Cette définition n'est valable que dans le cas du français. Chaque langue pouvant avoir un certain nombre de caractères propres.

-b) Chiffre

L'ensemble des chiffres noté CHIF, est donné par l'énumération suivante :

CHIF = { 0, 1, 2, 3, 4, 5, 6, 7, 8, 9 }

-c) Mot

Un mot M est une suite contiguë finie de caractères pris dans l'ensemble noté ALPH, constitué par l'union de l'ensemble des lettres et de l'ensemble des chiffres. On dira également mot simple, par opposition à mot composé.

$$ALPH = LET \cup CHIF$$

-d) Séparateur

On appelle séparateur tous les autres caractères utilisables pour l'écriture du français. Cet ensemble noté SEP comprend :

- les signes de ponctuation,
- les caractères représentant des opérateurs : +, -, =, <, ≤, etc...,
- les signes orthographiques que sont "l'apostrophe" et "le trait d'union",
- le caractère "blanc",

- ainsi qu'un ensemble de symboles spéciaux, qui doit être nécessairement paramétrable.

-e) Séparateur non-impératif

On distingue parmi l'ensemble des séparateurs, un sous-ensemble qui sera constitué de ce que nous nommons des séparateurs non-impératifs. Cette distinction sera utilisée pour définir les mots composés. Cet ensemble, noté SNI, est composé du caractère blanc (noté "espace" ci-dessous), de l'apostrophe, et du trait d'union.

$$SNI = \{ \text{espace} , \text{ ' } , \text{ - } \}$$

Par opposition, tout autre séparateur est appelé **séparateur impératif.**

-f) Mot composé

Un mot composé noté MC, est une suite contiguë et finie de mots séparés les uns des autres par un séparateur non impératif.

Exemple :

" *aujourd'hui* " " *pomme de terre* " " *ci-dessus* "

Remarque 3 :

Tout mot composé MC, n'appartient pas forcément à une langue naturelle, même si les mots simples le composant appartiennent à cette langue.

Exemple :

" *l'élève* " " *viendra-t-il* "

Remarque 4 :

Les nombres constituent un ensemble de mots particuliers, dans la mesure où ils sont constitués d'une suite contiguë de chiffres, auxquels peuvent se mêler dans certaines conditions des points et une virgule (dans notre analyseur, la reconnaissance des nombres est effectuée par une procédure particulière).

Exemple :

" 1.986.456,123 "

Remarque 5 :

La distinction entre le trait d'union, l'opérateur "moins", et le tiret de césure, sera effectué dans notre analyseur par une procédure particulière.

-g) Forme

On appelle forme, notée F_i, où i désigne le rang de la forme dans le texte, un mot simple, un mot composé, ou encore un séparateur impératif.

-h) Unité lexicale

On appelle unité lexicale un ensemble permettant de regrouper les différentes formes d'un même mot simple ou composé, de même catégorie grammaticale, représentant un même concept. Ces différentes formes proviennent généralement des marques orthographiques que prend un mot lorsque les variables grammaticales associées à sa catégorie peuvent admettre des valeurs différentes : "*s* " pour le pluriel, "*e* " pour le féminin, désinences verbales, etc... Dans un petit nombre de cas il peut s'agir de mots admettant plusieurs orthographes (" *réécrire* " ou " *récrire* ").

Remarque 6 :

Une unité lexicale peut être réduite à une forme unique, dans le cas où un mot n'admet qu'une seule forme.

-i) Représentant d'une unité lexicale

On appelle représentant d'une unité lexicale la forme normalisée d'un mot qui sera :

- dans l'ordre en fonction de l'existence de la forme, la forme au masculin singulier, sinon la forme au masculin pluriel, sinon la forme au féminin singulier, sinon la forme au féminin pluriel,
- la forme à l'infinitif dans le cas d'un verbe,
- la forme unique.

3.2. L'analyse morphologique

3.2.1. Solution morphologique et ensemble solution d'une forme

On appelle ensemble solution d'une forme F_i, noté SM_i, l'ensemble des triplets numérotés (représentant d'unité lexicale, catégorie grammaticale, ensemble des valeurs grammaticales), notés sm_{in}, qui représentent les interprétations syntaxiques hors-contexte dans le modèle linguistique utilisé. Nous appelons solutions morphologiques, les éléments $sm_{i,n}$ de l'ensemble solution SM_i. Cet ensemble SM_i est indépendant de tout processus d'analyse. Il constitue en fait, l'ensemble solution intrinsèque dans le modèle linguistique.

Exemple :

Pour F_i = "*cours* " on a :

$$SM_i = \{ sm_{i,1} , sm_{i,2} , sm_{i,3} , sm_{i,4} \}$$

avec :

$sm_{i,1}$ = { "*cour* " SUBC (féminin, pluriel)}

$sm_{i,2}$ = { "*cours* " SUBC (masculin, {singulier, pluriel})}

$sm_{i,3}$ = { "*courir* " VBCJ (indicatif, présent, {1ère,2ème}, singulier)}

$sm_{i,4}$= { "*courir* " VBCJ (impératif, présent, 2ème, singulier)}

Les deux solutions morphologiques $sm_{i,1}$ et $sm_{i,2}$ représentent deux interprétations possibles de la forme nominale "*cours* " qui différent par les valeurs des variables grammaticales genre et nombre, et qui correspondent à des concepts différents. Alors que les deux solutions morphologiques $sm_{i,3}$ et $sm_{i,4}$ représentent deux interprétations possibles de la forme verbale "*cours* " qui différent par les valeurs des variables grammaticales mode et personne, mais qui correspondent au même concept : "*courir* ".

En fait, chacune des deux solutions morphologiques $sm_{i,1}$ et $sm_{i,2}$ recouvre plusieurs concepts polysèmes que nous ne pouvons distinguer hors contexte (cf. l'exemple donné avec "*carte* " à la section précédente).

Une distinction entre deux solutions morphologiques comportant la même catégorie grammaticale pourra être utilisée lors du filtrage syntaxique pour les

vérifications des accords grammaticaux, lorsque les valeurs des variables grammaticales seront significatives.

Solution morphologique englobante

On dira qu'une solution morphologique sm_1 est englobante par rapport à une solution sm_2 comportant la même catégorie syntaxique si les valeurs grammaticales de sm_1 contiennent, au sens ensembliste, celles de sm_2.

3.2.2. Fonction de l'analyse morphologique

La fonction de l'analyse morphologique est de segmenter en formes F_i un texte écrit en langue naturelle, et de les interpréter par rapport au modèle linguistique utilisé, en leur attribuant un ensemble solution noté S_i de solutions morphologiques $s_{i,n}$ numérotées.

Cet ensemble solution est déterminé à l'aide des données linguistiques du système d'analyse : il est déterminé par le processus d'analyse morphologique, et représente les interprétations syntaxiques hors-contexte dans le système d'analyse, en fonction des données linguistiques utilisées (ces solutions peuvent donc être différentes des solutions hors-contexte intrinsèques dans le modèle linguistique).

Cette possible différence entre les deux ensembles solutions, S_i qui est déterminé par l'analyse morphologique et SM_i qui est indépendant de tout processus d'analyse, provient de la difficulté de mise en oeuvre d'un modèle linguistique. Cette difficulté est liée essentiellement à l'impossibilité pratique d'une prise en compte exhaustive au sein du système d'analyse des données linguistiques nécessaires.

Exemple :

si l'analyse morphologique a fourni p solutions morphologiques pour la forme F_i et que l'on note $s_{i,k}$ la $k^{ème}$, alors :

$$S_i = \{ s_{i,1}, \ldots, s_{i,k}, \ldots, s_{i,p} \}$$

Une analyse morphologique parfaite doit fournir, quelle que soit la forme analysée, un ensemble solution identique à l'ensemble solution hors-contexte dans le modèle linguistique. c'est-à-dire que :

$$\forall \; F_i \quad \text{on ait} \quad S_i = SM_i$$

Il est bien évident qu'il est utopique d'envisager une telle analyse, dès que le modèle morphologique utilisé est un tant soit peu discriminant, notamment pour les classes ouvertes, et que l'on appréhende des univers textuels ouverts (où le vocabulaire connu est un sous-ensemble du vocabulaire de la langue traitée). Notre souci sera de nous approcher le plus possible d'une telle analyse.

Une analyse morphologique va s'effectuer à partir d'un certain nombre de renseignements linguistiques initiaux définis en fonction du modèle linguistique utilisé, qui constituent les données linguistiques du système. Nous verrons par la suite qu'il est important de bien définir ces données linguistiques, pour obtenir une analyse morphologique optimale, et en particulier pour pouvoir inférer automatiquement un certain nombre de ces renseignements pour les mots inconnus du système.

Les deux problèmes essentiels qui se posent lors de l'analyse morphologique sont :

- d'une part, le cas où il n'y a pas de solution morphologique unique pour une forme,

- et d'autre part, le cas où l'analyseur manque d'informations pour traiter une forme;

ce qui nous amène au deux définitions suivantes :

-a) Forme ambiguë

On dira qu'une forme F_i est ambiguë, si et seulement si l'ensemble solution S_i qui lui est associé, comprend plusieurs éléments. Une forme ambiguë est aussi appelée traditionnellement forme homographe, et constitue une ambiguïté grammaticale.

F_i ambiguë $<==> | S_i | > 1$ où $| S_i |$ désigne le cardinal de S_i.

-b) Forme inconnue

On dira qu'une forme F_i est inconnue (du système d'analyse morphologique), si et seulement si l'ensemble solution S_i qui lui est associé est vide.

$$F_i \text{ inconnue } <==> |S_i| = 0 <==> S_i = \{\ \}$$

Remarque 7 :

Une forme peut être inconnue pour deux raisons :

- soit la forme n'appartient pas à la langue (faute d'orthographe, faute de frappe, mot créé dans le texte, etc...),

- soit la forme appartient à la langue et sa non-reconnaissance provient d'une insuffisance de l'analyse morphologique (données linguistiques incomplètes ou erronées, mauvaise segmentation).

Nous nous sommes surtout intéressé au deuxième cas envisagé : l'insuffisance de l'analyse morphologique, sans pour cela négliger l'importance du premier et notamment les fautes de frappe.

Les cas d'insuffisance de l'analyse morphologique résultant d'une mauvaise segmentation peuvent être résolus par une vérification, en faisant fonctionner l'analyseur morphologique en génération. Bien que fastidieuse cette vérification est nécessaire et constitue un aspect de la mise au point de l'analyseur.

Pour le cas d'une insuffisance de l'analyse morphologique due à des données linguistiques incomplètes, nous avons développé une stratégie de traitement des mots inconnus, nous permettant d'inférer pour une forme inconnue F_i un ensemble solution S_i, incluant l'ensemble solution hors contexte dans le modèle linguistique SM_i.

C'est ce que nous appelons abusivement attribution automatique de catégories grammaticales potentielles, puisque ce traitement basé sur notre distinction entre classes ouvertes et fermées permet également de déterminer un ensemble de valeurs grammaticales et d'affecter automatiquement une unité lexicale.

Nous détaillerons ce mécanisme au chapitre suivant (cf. V.6), et nous verrons en particulier comment on peut minorer le nombre d'éléments de l'ensemble solution obtenu.

La figure ci-après permet d'en schématiser la fonction.

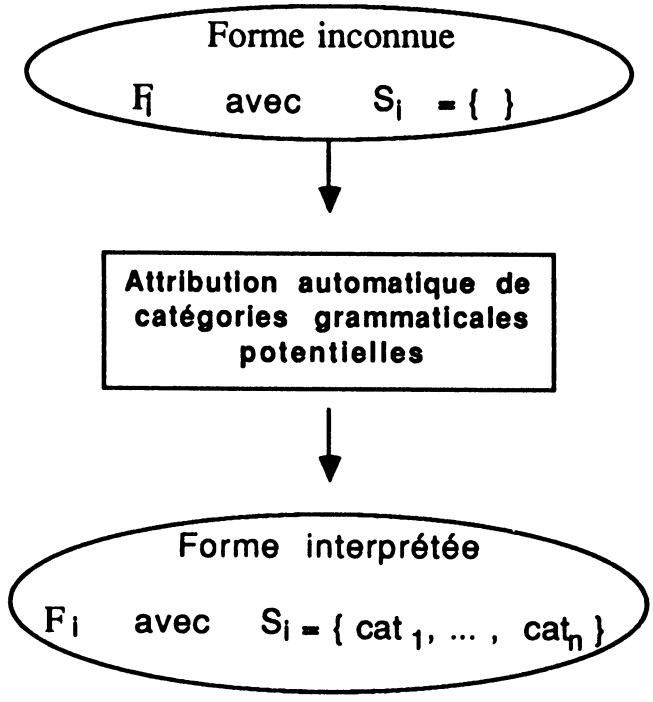


Figure 1 : *Attribution automatique de catégories grammaticales potentielles*

Ainsi à l'issue de notre analyse morphologique, il n'y aura plus de forme inconnue et un texte sera entièrement segmenté en formes interprétées (possédant un ensemble solution non vide).

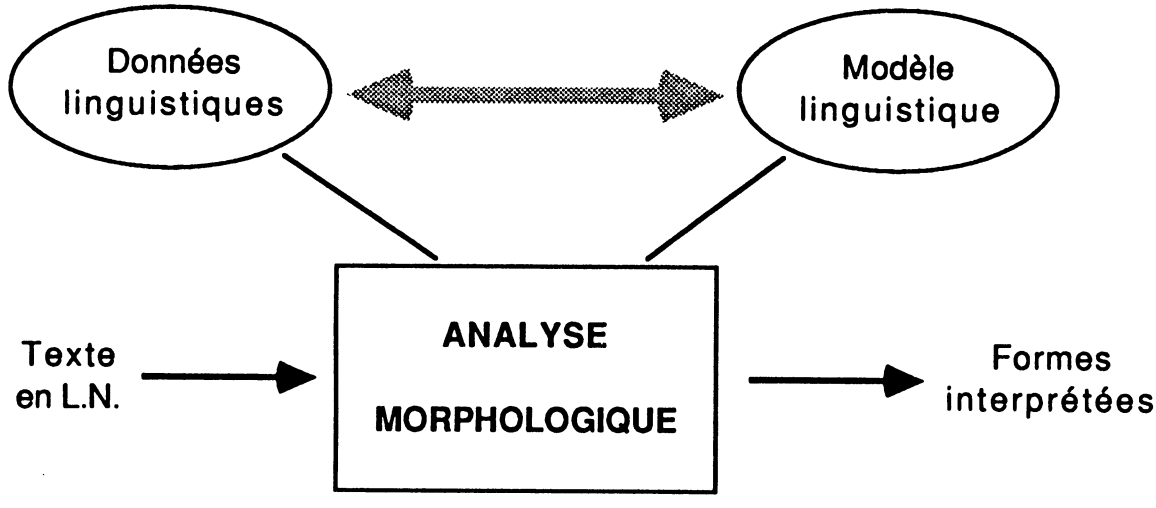


Figure 2 : *Analyse morphologique*

3.3. Le filtrage syntaxique

3.3.1. Introduction

Pour un texte écrit en langue naturelle, à partir du seul résultat de l'analyse morphologique, on peut obtenir différentes interprétations. Chaque interprétation correspondant au choix d'une solution morphologique pour chaque forme ambiguë du texte.

Exemple :

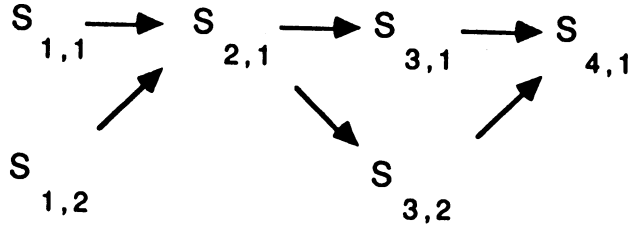
Pour un texte constitué de 4 formes, numérotées de 1 à 4, dont les ensembles solutions respectifs sont :

$S_1 = \{s_{1,1}, s_{1,2}\}$, $S_2 = \{s_{2,1}\}$, $S_3 = \{s_{3,1}, s_{3,2}\}$, $S_4 = \{s_{4,1}\}$

on a les quatre interprétations possibles, représentées par les succession des solutions morphologiques suivantes :

$s_{1,1}$	$s_{2,1}$	$s_{3,1}$	$s_{4,1}$
$s_{1,1}$	$s_{2,1}$	$s_{3,2}$	$s_{4,1}$
$s_{1,2}$	$s_{2,1}$	$s_{3,1}$	$s_{4,1}$
$s_{1,2}$	$s_{2,1}$	$s_{3,2}$	$s_{4,1}$

On représentera graphiquement par un arc la succession de deux solutions morphologiques appartenant aux ensembles solutions de deux formes consécutives. Pour l'exemple précédent on obtient le graphe suivant :



Or dans une langue naturelle, toutes les successions de solutions morphologiques ne sont pas forcément possibles. L'ordonnancement des mots est régi par un ensemble de règles formant la grammaire de la langue (même

s'il est difficile de définir complètement une grammaire pour une langue) portant, entre autres, sur les successions de catégories grammaticales et sur les accords grammaticaux (accords en genre, nombre, personne). C'est-à-dire que dans le graphe des solutions morphologiques, tous les arcs ne seront pas forcément valides.

3.3.2. Définition

On appelle **filtrage syntaxique**, un mécanisme permettant d'éliminer un certain nombre de successions illicites, c'est-à-dire, permettant d'éliminer un sous ensemble d'arcs dans le graphe potentiel établi à partir du résultat de l'analyse morphologique.

Ce filtrage syntaxique utilise les restrictions de la grammaire de la langue portant sur les successions de catégories grammaticales et sur les contraintes d'accord grammatical entre certaines classes syntaxiques. On peut définir ce filtrage en disant qu'il s'agit d'une **analyse grammaticale simple.**

La terminologie d'analyse grammaticale, désigne classiquement un processus de résolution d'ambiguïtés grammaticales résultant de l'analyse morphologique (on dit aussi ambiguïtés morphologiques). Une ambiguïté grammaticale est une catégorisation multiple d'une forme. Une exploration du proche contexte de la forme ambiguë permet dans bien des cas la résolution de ce problème.

Afin d'expliciter la notion de filtrage syntaxique, nous sommes amenés à donner les définitions suivantes :

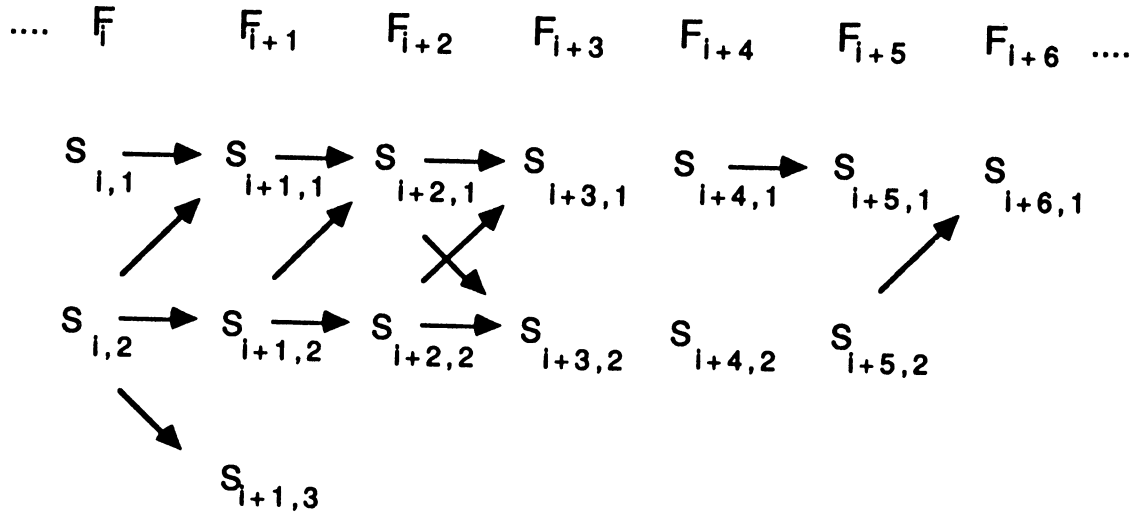
Transition

On appelle transition T_i d'une forme F_i à une forme F_{i+1}, l'ensemble des arcs reliant les solutions morphologiques de la solution S_i de F_i aux éléments de la solution S_{i+1} de F_{i+1}.

Si l'on note $s_{i,m}\, s_{i+1,n}$ l'arc reliant la $m^{ème}$ solution morphologique de S_i à la $n^{ème}$ de S_{i+1}, alors :

$$T_i = \{\, s_{i,m}\, s_{i+1,n} \mid S_i \ni s_{i,m}\,,\ S_{i+1} \ni s_{i+1,n} \,\}$$

Soient la suite de formes suivante et le graphe de solutions morphologiques associées, qui nous servira dans la suite de cet exposé à illustrer nos propos (il ne s'agit pas d'un graphe résultant de l'analyse morphologique qui serait alors complet) :



graphe 1

Exemple :

Sur le graphe 1 on a les transitions suivantes :

$$T_i = \{ s_{i,1}\, s_{i+1,1} \ , \ s_{i,2}\, s_{i+1,1} \ , \ s_{i,2}\, s_{i+1,2} \ , \ s_{i,2}\, s_{i+1,3} \}$$
$$T_{i+3} = \{ \ \} \quad \text{et} \quad T_{i+4} = \{ \ s_{i+4,1}\, s_{i+5,1} \ \}$$

On peut différencier quatre types de transitions qui sont : les transitions vides, les transitions incohérentes gauches ou droites, les transitions cohérentes, et les transitions complètes. Cette distinction permet de caractériser les facteurs de discontinuité dans l'interprétation d'une portion de texte.

-a) Transition vide

On dira qu'une transition T_i d'une forme F_i à une forme F_{i+1} est une transition vide, si et seulement si T_i est l'ensemble vide, c'est-à-dire s'il n'existe aucun arc reliant les solutions morphologiques de l'ensemble solution S_i de F_i à celles de l'ensemble solution S_{i+1} de F_{i+1}.

Exemple :

Sur le graphe 1 : T_{i+3} est une transition vide

-b) Transition incohérente gauche (droite)

On dira qu'une transition T_i d'une forme F_i à une forme F_{i+1} est incohérente gauche par rapport à la transition précédente T_{i-1} si et seulement si il n'existe pas de solution morphologique de la solution S_i, qui soit extrémité d'un arc de T_{i-1}, et qui soit origine d'un arc de T_i.

Exemple :

Sur le graphe 1 : T_{i+5} est incohérente gauche

De même on dira qu'une transition T_i d'une forme F_i à une forme F_{i+1} est incohérente droite par rapport à la transition suivante T_{i+1}, si et seulement si il n'existe pas de solution morphologique de la solution S_{i+1}, qui soit extrémité d'un arc de T_i, et qui soit origine d'un arc de T_{i+1}.

Exemple :

Sur le graphe 1 : T_{i+4} est incohérente droite

Par extension, on pourra dire qu'une forme F_i est incohérente gauche ou incohérente droite, si la transition T_i associée à cette forme, est incohérente gauche ou incohérente droite.

Exemple :

Sur le graphe 1 : F_{i+4} est incohérente droite et F_{i+5} est incohérente gauche

Propriété 1 :

Une transition vide est incohérente gauche et incohérente droite (la réciproque n'étant pas vraie).

Exemple :

Sur le graphe 1 :

T_{i+3} est une transition vide, incohérente gauche et droite,

T_{i+4} est incohérente gauche et droite, mais n'est pas vide

Propriété 2 :

Si une transition T_i est incohérente droite, alors la transition T_{i+1} est incohérente gauche (la réciproque est vraie) :

Exemple :

Sur le graphe 1 : T_{i+2} ,T_{i+3} ,T_{i+4} sont incohérentes droite,
T_{i+3} ,T_{i+4} ,T_{i+5} sont incohérentes gauche.

-c) Transition cohérente

On dira qu'une transition T_i est cohérente, lorsqu'elle n'est ni incohérente gauche, ni incohérente droite.

Exemple :

Sur le graphe 1 : T_{i+1} est cohérente

Par simplification on dira qu'une transition est incohérente lorsqu'elle est incohérente gauche ou incohérente droite.

Exemple :

Sur le graphe 1 : T_{i+2} , T_{i+3} , T_{i+4}, et T_{i+5} sont incohérentes

Par analogie, on parlera également de forme cohérente ou de forme incohérente. Il faudra entendre par là une forme dont l'analyse est cohérente ou incohérente dans un contexte donné.

-d) Transition complète

On dira qu'une transition T_i d'une forme F_i à une forme F_{i+1} est une transition complète, si et seulement si pour toute solution morphologique $s_{i,n}$ de la solution S_i de F_i et pour tout $s_{i+1,m}$ élément de la solution S_{i+1} de F_{i+1}, il existe un arc reliant $s_{i,n}$ à $s_{i+1,m}$.

Exemple :

Sur le graphe 1 : T_{i+2} est une transition complète

En employant cette terminologie, on peut dire que pour le graphe des solutions morphologiques obtenu après l'analyse morphologique et avant le filtrage syntaxique, toutes les transitions entre deux formes consécutives sont

potentiellement complètes (et donc cohérentes), car aucune sélection positionnelle n'a encore été effectuée.

3.3.3. Fonction du filtrage syntaxique

La fonction du filtrage syntaxique est donc d'examiner les transitions une à une et d'éliminer les arcs illicites liant les solutions morphologiques. Pour certaines successions seules seront prises en compte les catégories grammaticales, pour d'autres les valeurs de certaines variables grammaticales interviendront. Ce filtrage syntaxique est réalisé par l'exploitation d'une matrice de précédence multivaluée que nous présentons au chapitre suivant (cf. V.4.3).

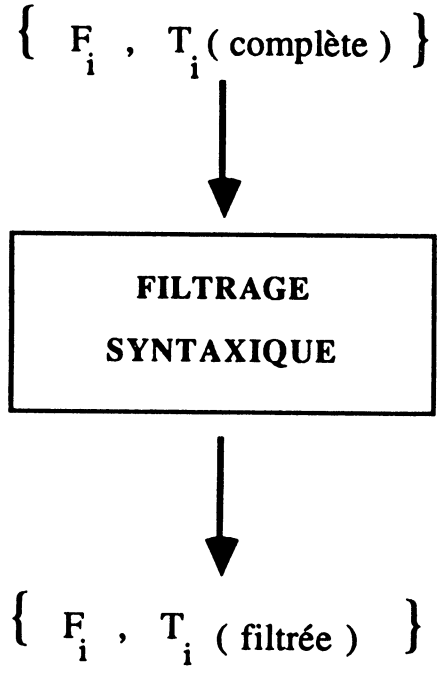


Figure 3 : *Filtrage syntaxique*

Une portion de texte comprise entre deux formes F_i et F_j peut avoir à l'issue de l'analyse morphologique et du filtrage syntaxique plusieurs interprétations, si elle comporte au moins une forme ambiguë (et que toutes les formes sont cohérentes). Nous allons voir maintenant comment interpréter le résultat obtenu après ce filtrage syntaxique.

3.4. Interprétation d'une portion de texte

Nous allons formaliser cette notion d'interprétation d'une portion de texte, en introduisant les notions de chemin et de parcours entre deux formes. Un chemin correspondant à une interprétation particulière d'une portion de texte, un parcours correspondant à l'ensemble des interprétations possibles pour une portion de texte donnée.

3.5. Notions de chemin

1ère définition :

On appelle chemin $C_{i,j}$ de longueur $k = j - i$ (avec $i < j$) entre deux formes F_i et F_j, tout ensemble de k arcs permettant de passer d'un élément de la solution S_i à un élément de la solution S_j, constituant une chaîne de $T_i, T_{i+1},...,T_{j-1}$:

$$\{ s_{i,m}\, s_{i+1,n} , s_{i+1,n}\, s_{i+2,p} , \ldots , s_{j-1,q}\, s_{j,r} \}$$

Exemple :

Sur le graphe 1 :

$\{s_{i,2}s_{i+1,1} , s_{i+1,1}s_{i+2,1} , s_{i+2,1}s_{i+3,2}\}$ est un chemin de F_i à F_{i+3}.

Afin de simplifier l'écriture d'un chemin, on introduit une deuxième définition, équivalente à la première, formulée sur l'ensemble des sommets appartenant au chemin.

2ème définition :

On appelle chemin $C_{i,j}$ de longueur $k = j - i$ (avec $i < j$) entre 2 formes F_i et F_j, un ensemble de k+1 solutions morphologiques numérotées de i à j avec $j = i + k$:

Tout ensemble de solutions morphologiques :

$$S_i \times S_{i+1} \times \ldots \times S_j \ni \{ s_{i,m} , s_{i+1,n} , \ldots , s_{j,q} \}$$

où $\forall$ p, p=i,...,j, on ait $S_p \ni s_{p,n}$

et $\forall$ p, p=i,...,j-1 on ait $T_p \ni s_{p,n}\, s_{p+1,m}$

est un chemin $C_{i,j}$

Exemple :

Sur le graphe 1 :

$\{s_{i,2}, s_{i+1,1}, s_{i+2,1}, s_{i+3,2}\}$ est un chemin liant F_i à F_{i+3}.

Propriété 3 :

Le nombre maximal de chemins potentiels entre 2 formes F_i et F_j, est donné par le cardinal du produit cartésien des ensembles solutions S_p pour p=i,...,j, soit la formule :

$$\prod_{p=i,j} |S_p| \quad \text{où } |S_p| \text{ désigne le cardinal de l'ensemble } S_p.$$

Pour une portion de texte comprise entre deux formes F_i et F_j, avant le filtrage syntaxique, le nombre maximal d'interprétations possibles est donné par cette formule.

Propriété 4 :

On dira qu'il n'existe pas de chemin entre deux formes F_i et F_j (avec $j - i \leq 2$), si et seulement si il existe une forme F_k, avec $i \leq k < j-1$, pour laquelle la transition associée T_k est incohérente droite (ou respectivement une forme F_{k+1}, pour laquelle T_{k+1} est incohérente gauche)

Exemple :

Sur le graphe 1 : Il n'existe pas de chemin entre F_i et F_{i+4}, car T_{i+2} est incohérente droite.

Propriété 4bis :

Il n'existe pas de chemin entre deux formes consécutives F_i et F_{i+1} si et seulement si la transition T_i est vide.

Il n'existe pas de chemin entre F_i et F_j <==> $T_i = \{\ \}$

Exemple :

Sur le graphe 1 : Il n'existe pas de chemin entre F_{i+3} et F_{i+4}.

Chemin pendant

On appelle chemin pendant noté CP, entre deux formes F_i et F_j un chemin tel que l'une des deux conditions suivantes soit remplie :

- soit l'extrémité du chemin est une solution morphologique appartenant à l'ensemble solution S_j, telle qu'il n'existe aucun arc dans la transition T_j ayant cette solution morphologique comme origine,
- soit, l'origine du chemin est une solution morphologique appartenant à l'ensemble solution S_i, telle qu'il n'existe aucun arc dans la transition T_{i-1} ayant cette solution morphologique comme extrémité.

Exemples :

Sur le graphe 4 (défini plus loin) :

seul le chemin $\{ s_{i+3,1}, s_{i+4,2}, s_{i+5,2} \}$ est un chemin pendant.

Ces chemins pendants constituent des impasses pour les tentatives d'interprétation d'une portion de texte; ils représentent en quelque sorte des embryons d'interprétation incohérente. Pour compléter cette notion d'interprétation incohérente, on appellera solution morphologique isolée, une solution morphologique d'une forme F_i qui n'est origine ou extrémité d'aucun arc des transitions T_i et T_{i-1}.

Exemple :

Sur le graphe 1 : $s_{i+4,2}$ est une solution isolée.

Remarque 8 :

S'il n'existe pas de chemin entre deux formes F_i et F_j , cela signifie qu'on ne peut pas avoir, à partir de l'analyse effectuée, une interprétation cohérente de la portion de texte comprise entre ces deux formes. Il peut y avoir deux explications à ce phénomène :

- le texte écrit est incorrect,
- il existe au moins une forme de ce texte, pour laquelle l'analyse est incomplète, c'est-à-dire que l'on n'a pas obtenu toutes les interprétations possibles pour cette forme.

Nous verrons par la suite comment exploiter cette remarque pour tenter de compléter automatiquement une telle analyse, afin d'obtenir les interprétations "oubliées" (cf. V.6.6).

3.5.1. Notions de parcours

3.5.1.1. Définition

Parcours

On appelle parcours $P_{i,j}$ d'une forme origine F_i à une forme extrémité F_j (défini pour $i < j$) l'ensemble des chemins menant de F_i à F_j.

Exemple :

Sur le graphe 1 : Le parcours $P_{i,i+3}$ est composé des huit chemins suivants :

$\{ s_{i,1} , s_{i+1,1} , s_{i+2,1} , s_{i+3,1} \}$
$\{ s_{i,1} , s_{i+1,1} , s_{i+2,1} , s_{i+3,2} \}$
$\{ s_{i,2} , s_{i+1,1} , s_{i+2,1} , s_{i+3,1} \}$
$\{ s_{i,2} , s_{i+1,1} , s_{i+2,1} , s_{i+3,2} \}$
$\{ s_{i,2} , s_{i+1,2} , s_{i+2,1} , s_{i+3,1} \}$
$\{ s_{i,2} , s_{i+1,2} , s_{i+2,1} , s_{i+3,2} \}$
$\{ s_{i,2} , s_{i+1,2} , s_{i+2,2} , s_{i+3,1} \}$
$\{ s_{i,2} , s_{i+1,2} , s_{i+2,2} , s_{i+3,2} \}$

Propriété 5 :

On dira que la longueur d'un parcours $P_{i,j}$ est la longueur $k=j-i$ commune aux chemins menant de F_i à F_j.

Exemple :

Sur le graphe 1 : la longueur du parcours $P_{i,i+3}$ est : $k = 3$

Propriété 6 :

Le cardinal d'un parcours $P_{i,j}$ est le nombre de chemins (tous de longueur $k = j - i$) qui le composent.

Exemple :

Sur le graphe 1 : le cardinal du parcours $P_{i,i+3}$ est : $| P_{i,i+3} | = 8$

Le cardinal d'un parcours entre deux formes F_i et F_j , correspond au nombre d'interprétations possibles de la portion de texte comprise entre ces deux formes. Nous avons vu (cf. propriété 3) que ce nombre est borné par le produit cartésien des ensembles solutions : S_i X ....X S_j

Propriété 7 :

On dira qu'un parcours $P_{m,n}$ est enchâssé dans un parcours $P_{i,j}$ si et seulement si $i \leq m < n \leq j$.

Exemple :

Sur le graphe 1 :

$P_{i,i+1}$, $P_{i,i+2}$, $P_{i,i+3}$, $P_{i+1,i+2}$, $P_{i+1,i+3}$, $P_{i+2,i+3}$,
sont tous des parcours enchâssés dans le parcours $P_{i,i+3}$.

3.5.1.2. Notion de parcours impossible

Parcours impossible

On dira qu'un parcours $P_{i,j}$ est impossible si et seulement si il n'existe aucun chemin menant de F_i à F_j.

$$P_{i,j} \text{ impossible } <==> |P_{i,j}| = 0 \quad (P_{i,j} = \{\ \})$$

Exemple :

Sur le graphe 1 : $P_{i,i+4}$ est impossible

Propriété 8 :

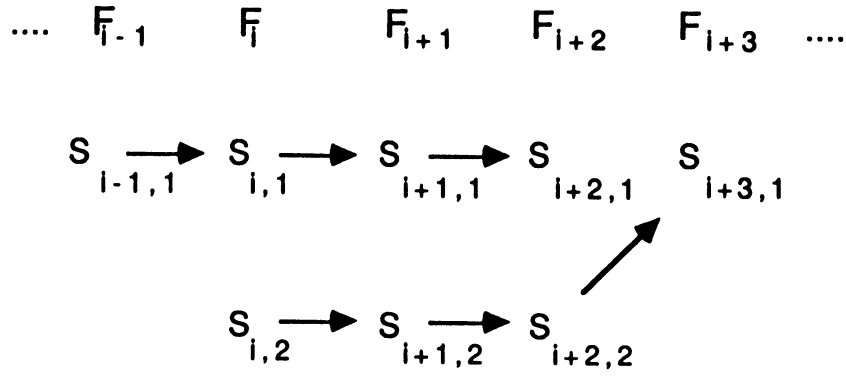
Une condition suffisante, mais non nécessaire, pour qu'un parcours $P_{i,j}$ soit impossible, est qu'il existe au moins une transition T_k pour $i \leq k < j-1$, qui est incohérente droite (ou respectivement T_{k+1} , incohérente gauche) :

$$\exists\, k,\ i \leq k < j-1 \ | T_k \text{ soit incohérente droite } ==> P_{i,j} \text{ impossible}$$

Exemple :

Sur le graphe 1 : $P_{i,i+4}$ est impossible , T_{i+2} est incohérente droite.

Sur le graphe 2, le parcours $P_{i,i+3}$ est impossible alors que les transitions T_i , T_{i+1} et T_{i+2} sont cohérentes.



graphe 2

3.5.1.3. Notions de Parcours ambigu et de parcours linéaire

Parcours ambigu

On dira qu'un parcours $P_{i,j}$ est ambigu si et seulement si il contient au moins deux éléments (il existe au moins 2 chemins menant de F_i à F_j).

$$P_{i,j} \text{ ambigu } <==> |P_{i,j}| > 1$$

Exemple :

Sur le graphe 2 : $P_{i,i+2}$ est ambigu

Si un parcours entre deux formes F_i et F_j est ambigu, cela signifie que la portion de texte comprise entre ces deux formes est ambiguë (c.a.d. qu'elle comporte plusieurs interprétations possibles).

Parcours linéaire

On dira qu'un parcours $P_{i,j}$ est linéaire si et seulement si il n'existe qu'un seul chemin menant de F_i à F_j, soit :

$$P_{i,j} \text{ linéaire } <==> |P_{i,j}| = 1$$

Exemple :

Sur le graphe 1 : seuls les parcours $P_{i+4,i+5}$ et $P_{i+5,i+6}$ sont linéaires.

Si un parcours entre deux formes F_i et F_j est linéaire, cela signifie que la portion de texte comprise entre ces deux formes n'est pas ambiguë (une seule interprétation possible), en l'état des connaissances du système.

3.6. Résolution des ambiguïtés grammaticales

Pour terminer cette présentation, nous allons introduire la notion de réseau, qui va nous permettre de formaliser ce que nous avons jusqu'à maintenant appelé graphe de solutions morphologiques. La notion de réseau étant classique, nous allons nous contenter de l'exprimer à partir de la terminologie que nous venons de définir.

C'est cette notion de réseau qui va nous permettre d'introduire le processus de résolution des ambiguïtés grammaticales.

3.6.1. Notions de réseau

3.6.1.1. Définition

On appelle réseau $R_{i,j}$ de forme origine F_i et de forme extrémité F_j, le graphe m-parti ($m = j - i + 1$),
dont les noeuds sont donnés par $S_i \cup ... \cup S_j$,
et dont les arcs sont donnés par $T_i \cup ... \cup T_{j-1}$.

Un graphe m-parti est de la forme suivante :

l'ensemble des sommets est partitionné en m sous-ensembles, et il n'y a pas d'arc liant deux éléments d'un même ensemble.

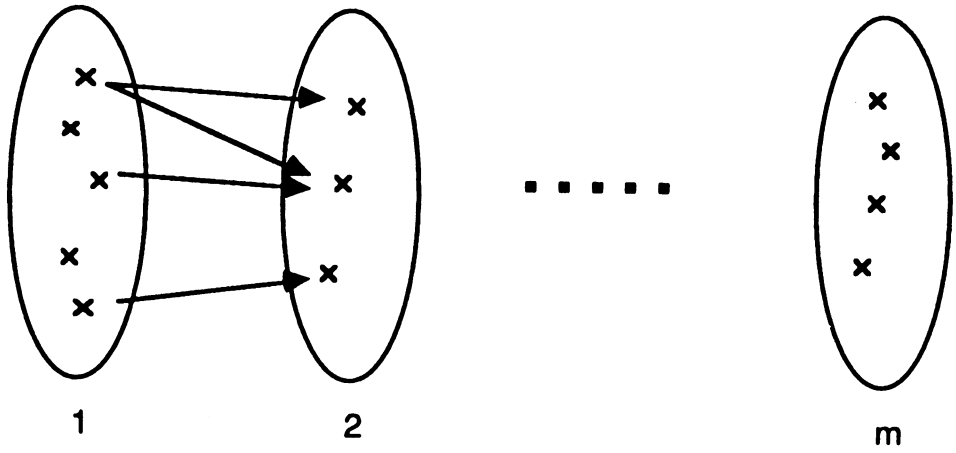


Figure 4 : *Graphe m-parti*

Propriété 9 :

La longueur d'un réseau $R_{i,j}$ est égale au nombre $k = j - i$ de transitions qu'il contient.

Remarque 9 :

Un réseau contient potentiellement toutes les interprétations possibles de la portion de texte sur laquelle il est défini, ainsi que tous les embryons d'interprétation incomplète (chemins pendants et solutions morphologiques isolées).

Dans la suite du traitement, nous nous intéresserons à deux types particuliers de réseaux :

- les réseaux linéaires (représentant une interprétation unique d'une portion de texte),
- et les réseaux ambigus (représentant une ambiguïté d'interprétation d'une portion de texte).

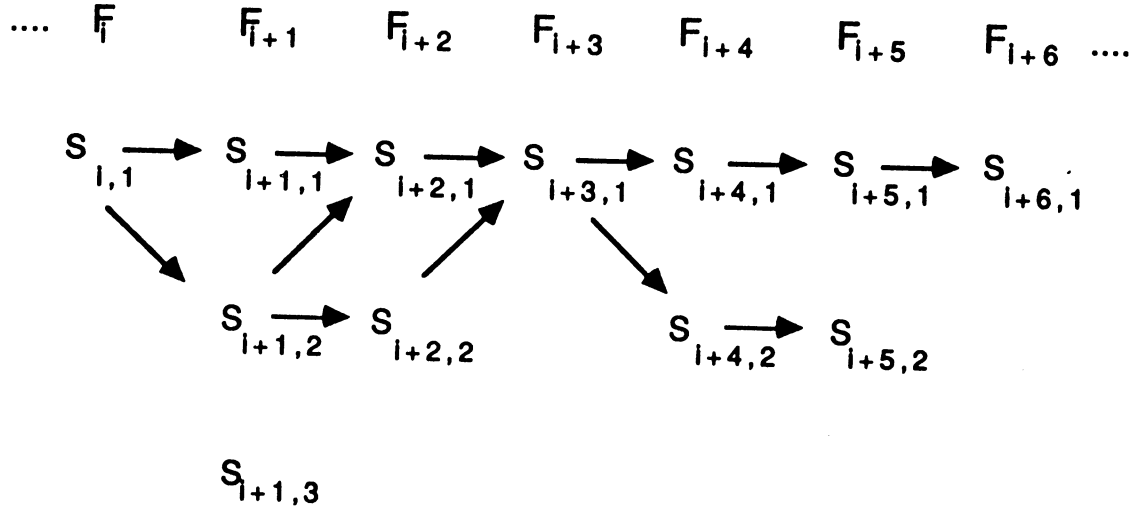
3.6.1.2. Notions de réseau ambigu et de réseau linéaire

Réseau ambigu

On dira qu'un réseau noté $RA_{i,j}$, est ambigu, si et seulement si le parcours $P_{i,j}$ est ambigu :

$$RA_{i,j} \text{ ambigu} \iff P_{i,j} \text{ ambigu} \iff |P_{i,j}| > 1$$

Soient la suite de formes suivante et le réseau associé :



graphe 3

Exemple :

Sur le graphe 3 :

le réseau $R_{i,i+3}$ est un réseau ambigu ($|P_{i,i+3}| = 3$)

le réseau $R_{i+3,i+5}$ est également ambigu ($|P_{i+3,i+5}| = 2$)

Réseau linéaire

On dira qu'un réseau noté $RL_{i,j}$, est linéaire, si et seulement si le parcours $P_{i,j}$ est linéaire. C'est-à-dire, si et seulement si il n'existe qu'un seul chemin menant de la forme origine F_i à la forme extrémité F_j :

$$RL_{i,j} \text{ linéaire} \iff P_{i,j} \text{ linéaire} \iff |P_{i,j}| = 1$$

Exemple :

Sur le graphe 3 : le réseau $R_{i+3,i+6}$ est linéaire.

Ces deux notions de réseau linéaire et de réseau ambigu sont essentielles pour la suite de l'exposé. Elles vont nous permettre de définir le processus de résolution des ambiguïtés et de spécifier le résultat fourni par l'analyseur de surface.

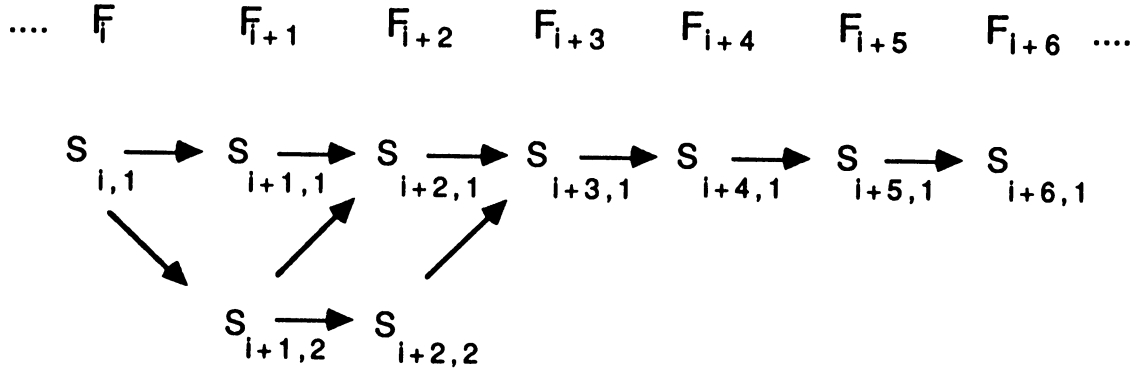
Avant cela, il est nécessaire d'introduire la notion de réseau simplifié. En effet nous avons vu précédemment qu'un réseau comporte tous les embryons d'interprétations incomplètes (cf. remarque 9), ce qui risque d'alourdir considérablement le processus d'analyse (exploration d'impasses). Or ces portions d'interprétations ne seront plus utilisées par la suite, il est donc plus simple de les supprimer.

Réseau simplifié

On appelle réseau simplifié noté $RS_{i,j}$ de forme origine F_i et de forme extrémité F_j , le réseau $R_{i,j}$ duquel on a ôté les chemins pendants et les solutions morphologiques isolées.

Exemple :

Si pour le réseau $R_{i,i+6}$ du graphe 3, on supprime le chemin pendant $\{s_{i+3,1}, s_{i+4,2}, s_{i+5,2}\}$ et la solution morphologique isolée $s_{i+1,3}$, on obtient le réseau simplifié $RS_{i,i+6}$ qui est représenté sur le graphe 4.



graphe 4

Propriété 10 :

On dira qu'un réseau noté $RM_{i,j}$, est linéaire maximum, si et seulement si :

- le réseau $R_{i,j}$ est linéaire,
- et que les réseaux simplifiés $RS_{i-1,j}$ et $RS_{i,j+1}$ si ils existent, sont ambigus.

Exemple :

Sur le graphe 4 : si l'on suppose qu'il n'existe pas de forme F_{i+7}, alors le réseau $R_{i+3,i+6}$ est linéaire maximum. En effet :

- le parcours $P_{i+3,i+6}$ est linéaire ($| P_{i+3,i+6} | = 1$)
- et le réseau simplifié $RS_{i+2,i+6}$ est ambigu ($| P_{i+2,i+6} | = 2$)

Remarque 10 :

On pourra toujours considérer un réseau $R_{i,j}$ comme la concaténation, notée + , d'une suite de réseaux linéaires maximums et de réseaux ambigus.

Exemple :

Sur le graphe 4 : on a $R_{i,i+6} = RA_{i,I+3} + RM_{i+3,i+6}$

Nous allons définir maintenant la notion de schéma de réseau, qui constitue la représentation formelle que nous utiliserons dorénavant pour manipuler ces objets.

3.6.2. Notions de schéma de transition et de réseau

3.6.2.1. Définition

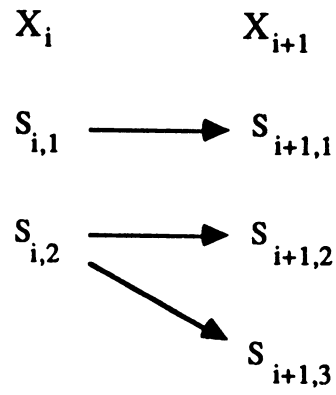
Schéma de transition

Par analogie à la définition d'une transition, on définit un schéma de transition noté ST, entre deux formes virtuelles X_i et X_{i+1}, en donnant un ensemble d'arcs dont les origines et les extrémités sont des solutions morphologiques (on dit que les deux formes X_i et X_{i+1} sont virtuelles, car elles ne correspondent pas à des formes trouvées

dans un texte, ce sont en fait des variables non instanciées qui représentent des formes).

Exemple :

$ST = \{ s_{i,1}\ s_{i+1,1}\ ,\ s_{i,2}\ s_{i+1,2}\ ,\ s_{i,2}\ s_{i+1,3} \}$ est un schéma de transition que l'on peut représenter graphiquement de la manière suivante :



Propriété 11 :

On dira qu'un schéma de transition ST est inclus dans une transition T_i si et seulement si chaque arc de ST est un arc de T_i (il s'agit de l'inclusion de graphes classique) :

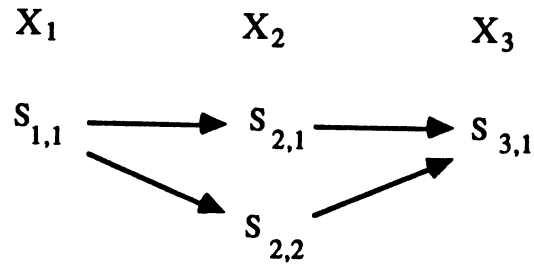
$$T_i \supseteq ST \iff \forall\ s_m\ s_n \mid ST \ni s_m\ s_n \quad \text{on ait} \quad T_i \ni s_m\ s_n$$

Schéma de réseau

On définit un schéma de réseau noté, SR^n, de longueur n, par la donnée d'un ensemble de schémas de transitions consécutives numérotés de 1 à n : $SR^n = \{ ST_1 , \ldots , ST_n \}$

Exemple :

$SR^2 = \{ \{ s_{1,1}\ s_{2,1}\ ,\ s_{1,1}\ s_{2,2} \}\ ,\ \{ s_{2,1}\ s_{3,1}\ ,\ s_{2,2}\ s_{3,1} \} \}$ est un schéma de réseau de longueur 2, que l'on peut représenter graphiquement de la manière suivante, avec X_1, X_2, X_3 qui sont des formes virtuelles :



Propriété 12 :

On dira qu'un schéma de réseau $SR^n = \{ ST_1, ..., ST_n \}$ est inclus dans un réseau $R_{i,j}$, si et seulement si le réseau $R_{i,j}$ comprend au moins n transitions consécutives $T_k, T_{k+1}, ..., T_{k+n-1}$, telles que :

$$T_k \supseteq ST_1, \quad T_{k+1} \supseteq ST_2, \quad ..., \quad T_{k+n-1} \supseteq ST_n$$

Exemple :

posons pour le schéma de réseau SR^2 défini dans l'exemple précédent :

$$s_{1,1} = s_{i,1} \ , \ s_{2,1} = s_{i+1,1} \ , \ s_{2,2} = s_{i+1,2} \ , \ s_{3,1} = s_{i+2,1}$$

alors le schéma de réseau SR^2 défini dans cet exemple , est inclus dans le réseau $R_{i,i+6}$ de le graphe 4; en effet :

$$T_i \supseteq ST_1 \text{ et } T_{i+1} \supseteq ST_2$$

3.6.2.2. Schémas linéaires et schémas ambigus

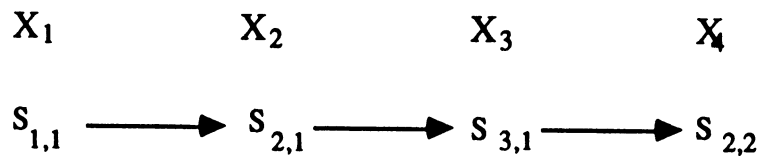
Par analogie aux notions de réseau linéaire et de réseau ambigu, on peut définir les notions de schéma linéaire et de schéma ambigu.

Schéma linéaire

On dira qu'un schéma de réseau SR^n de forme origine virtuelle X_1 et de forme extrémité virtuelle X_{n+1} est un schéma linéaire noté SL^n, si et seulement si le parcours $P_{1,n+1}$ défini entre X_1 et X_{n+1} est linéaire.

Exemple :

$SR^3 = \{ \{ s_{1,1}\ s_{2,1} \}, \{ s_{2,1}\ s_{3,1} \}, \{ s_{3,1}\ s_{4,1} \} \}$ est un schéma linéaire SL^3, que l'on peut représenter graphiquement de la manière suivante :



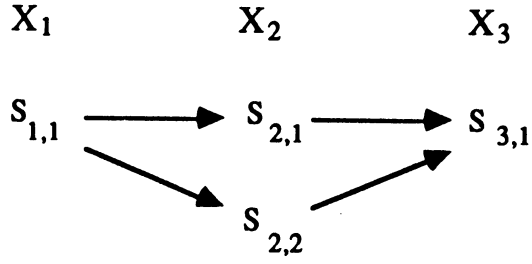
où X_1, X_2, X_3, X_4 sont des formes virtuelles, et le parcours $P_{1,4}$ est linéaire ($| P_{1,4} | = 1$).

Schéma ambigu

On dira qu'un schéma de réseau SR^n de forme origine virtuelle X_1 et de forme extrémité virtuelle X_{n+1} est un schéma ambigu noté SA^n, si et seulement si le parcours $P_{1,n+1}$ défini entre X_1 et X_{n+1} est ambigu.

Exemple :

Le schéma de réseau $SR^2 = \{ \{ s_{1,1}\ s_{2,1}\ ,\ s_{1,1}\ s_{2,2} \}, \{ s_{2,1}\ s_{3,1}\ ,\ s_{2,2}\ s_{3,1} \} \}$ est un schéma ambigu SA^2, car le parcours $P_{1,3}$ associé est ambigu : $| P_{1,3} | = 2$.



où X_1, X_2, X_3 sont des formes virtuelles.

Les schémas linéaires et les schémas ambigus vont nous servir à classifier respectivement les portions linéaires et les portions ambiguës d'un réseau. En effet un schéma, linéaire ou ambigu, établit un modèle dont l'inclusion pourra être testée à plusieurs endroits d'un même réseau.

Nous nous intéresserons par la suite à certains schémas de réseau linéaires, qui nous serviront à définir le transducteur d'états finis déterministe d'extraction des groupes conceptuels spécifié au chapitre précédent (cf. III.3), et à certains schémas ambigus, qui nous serviront à la résolution d'ambiguïtés grammaticales jugées intéressantes (cf. V.5).

4. Conclusion

On veut disposer d'un outil d'analyse permettant de transformer un texte écrit en langue naturelle en une forme traitable par un transducteur d'état fini déterministe dont le rôle sera la génération de structures particulières prédéfinies, en l'occurrence les groupes conceptuels (cf. III.2 et III.3).

Cette forme traitable par un transducteur d'état fini déterministe doit être un réseau linéaire simplifié, selon la terminologie précédemment définie (par la suite nous dirons réseau linéaire). Les groupes conceptuels étant des squelettes de syntagmes nominaux normalisés, la chaîne d'entrée du transducteur, devra être un réseau linéaire comprenant un syntagme nominal.

Le résultat de l'analyse grammaticale est un réseau de solutions morphologiques que l'on peut décomposer en une succession continue de réseaux linéaires maximums et de réseaux ambigus.

Si l'on veut, pour un texte donné, extraire le maximum de groupes conceptuels, il faut obtenir que le plus grand nombre de réseaux incluant un syntagme nominal soient linéaires. Pour cela on doit s'intéresser tout particulièrement à la linéarisation des réseaux ambigus susceptibles de contenir un groupe nominal.

C'est ainsi que nous nous sommes intéressés à une classification des ambiguïtés contenant des constituants potentiels de syntagmes nominaux, en utilisant des modèles définis au moyen de schémas ambigus. Seules les ambiguïtés débouchant systématiquement sur la même résolution seront retenues et modélisées, avec leur résolution, dans un catalogue de **schémas de résolution d'ambiguïté**. La construction de ce catalogue est fondée sur une étude systématique des ambiguïtés contenues dans les réseaux (cf. V.5.5).

L'analyse que nous effectuons est donc une analyse grammaticale partielle (morphologie + résolution de certaines ambiguïtés) dont l'objectif est la désambiguïsation des portions de texte susceptibles de contenir des syntagmes nominaux. Cette analyse se situe au niveau de la structure de surface de la langue naturelle, et fournit en sortie, une chaîne de formes catégorisées, dont les portions linéaires constitueront l'entrée du processus d'extraction des groupes conceptuel

La réalisation de cet outil d'analyse, qui constitue un analyseur de surface nécessite le développement d'un certain nombre de modules linguistiques qui sont :

-1) un analyseur morphologique s'appuyant sur un ensemble de constantes linguistiques déterminées en fonction du modèle linguistique défini;

-2) un processus de catégorisation automatique, permettant d'attribuer à une forme inconnue du système un ensemble de solutions morphologiques déterminées à partir des données linguistiques présentes dans le système;

-3) un processus de filtrage syntaxique s'appuyant sur les relations positionnelles et les contraintes d'accords grammaticaux de la langue (ces relations et ces accords constituent également des constantes linguistiques);

-4) un processus de résolution de certaines ambiguïtés grammaticales (cataloguées par des schémas de résolution d'ambiguïté) utiles pour la reconnaissance de syntagmes nominaux;

-5) un processus d'acquisition automatique du vocabulaire nouveau, permettant d'enrichir le dictionnaire d'analyse au fur et à mesure de son apparition. Ce processus est indispensable à l'appréhension d'univers textuels ouverts.

Ce sont ces différentes composantes de l'analyseur de surface que nous allons détailler dans le prochain chapitre.

CHAPITRE V

PLAN DU CHAPITRE V

L'ANALYSEUR DE SURFACE

L'ANALYSEUR DE SURFACE

1. Introduction

Nous allons présenter dans ce chapitre, l'architecture et le fonctionnement des différents modules composant l'analyseur de surface que nous venons de spécifier au chapitre précédent. Ces modules sont au nombre de cinq et réalisent les fonctions suivantes :

-1) l'analyse morphologique,

-2) le filtrage syntaxique (positionnel et grammatical),

-3) la résolution des ambiguïtés grammaticales reconnues à l'aide de schémas de résolution répertoriés,

-4) la catégorisation automatique (des formes inconnues, et des formes apprises incohérentes gauches ou droites), c'est-à-dire la détermination des catégories syntaxiques potentielles et des valeurs grammaticales associées,

-5) l'enrichissement automatique du vocabulaire (non ambigu).

Cette présentation séquentielle nous a paru nécessaire pour une meilleure appréhension du fonctionnement de l'analyseur, bien qu'en pratique, certains modules soient fortement imbriqués. L'imbrication de modules, d'essence morphologique comme la catégorisation automatique, et d'essence syntaxique comme le filtrage positionnel, permet une utilisation optimale des renseignements linguistiques disponibles à ces différents niveaux. Cette coopération est notamment utilisée lors du traitement des formes inconnues et lors du contrôle de cohérence grammaticale effectué pendant le filtrage syntaxique.

Dans cet exposé, nous insisterons sur la constitution des ensembles de données initiaux, nécessaires pour chacun des modules. Nous distinguerons en particulier ceux dont la constitution peut s'opérer plus ou moins automatiquement par apprentissage (exemple : relevé des relations positionnelles nécessaires au filtrage syntaxique, recensement des différents schémas de résolution d'ambiguïté, etc...), de ceux dont la nécessité d'exhaustivité requiert souvent un fastidieux travail manuel préliminaire (exemple : listes d'exceptions grammaticales, recherche des homographes des classes fermées, etc...).

Nous allons commencer par une description synthétique de l'analyseur, puis nous détaillerons l'enchaînement et le fonctionnement des modules qui le composent. Chaque étape sera illustrée à l'aide d'une phrase exemple, qui nous servira de fil conducteur tout au long de ce chapitre.

2. Description synthétique de l'analyseur

Le schéma de la figure 1 présente les différents traitements subis par le texte d'entrée en langue naturelle. Il permet de visualiser les outils exploitant les informations linguistiques du système, qui sont utilisés au cours des différents stades de l'analyse. Une présentation détaillée de leur architecture sera donnée lors de la description des modules correspondants.

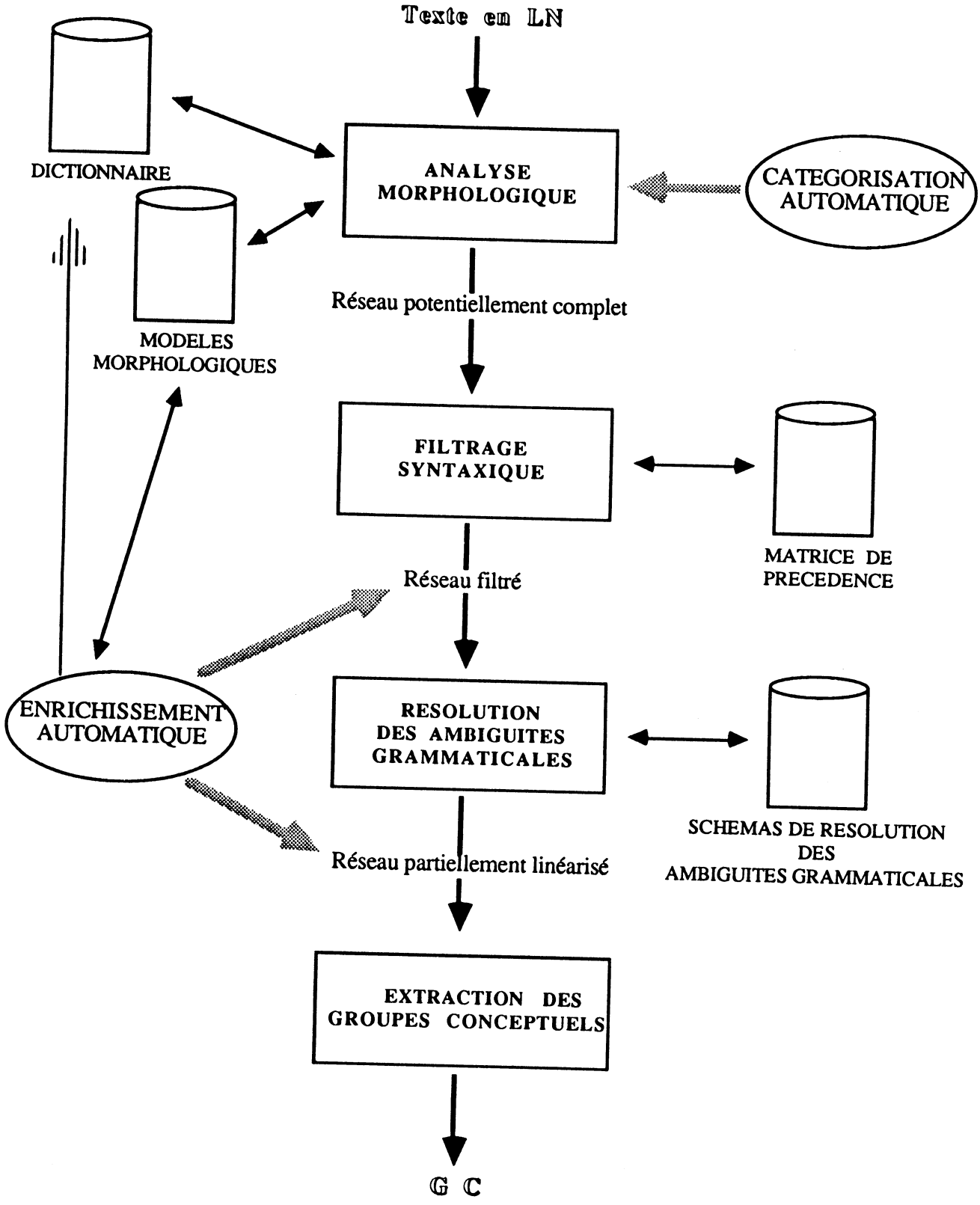


Figure 1 : *Schéma de fonctionnement*

Dans la pratique, les phases d'analyse morphologique, d'enrichissement automatique du vocabulaire et de traitements syntaxiques (filtrage avec la matrice de précédence et résolution d'ambiguïtés) sont fortement imbriquées. Le processus d'enrichissement automatique du vocabulaire par exemple, bien que d'essence morphologique, nécessite l'utilisation d'outils syntaxiques (filtrage) afin de déterminer la bonne catégorisation. L'imbrication de ces différents traitements est autorisée par la stratégie d'analyse sans retour arrière mise en oeuvre pour la morphologie. Cette stratégie permet de fournir pour une forme analysée, en une seule passe, les différentes catégorisations possibles. La structure du dictionnaire que nous présenterons plus avant (cf. 3.3.2.1.), autorise une analyse caractère par caractère, et permet de produire à partir d'une seule consultation les solutions morphologiques de chaque forme.

La première étape consiste en une analyse morphologique classique, segmentant le texte en formes reconnues à partir d'un dictionnaire de racines et d'un ensemble de désinences. Chaque décomposition génère une solution grammaticale. Lors de cette analyse le système mémorise un certain nombre de formes successivement analysées, de manière à pouvoir faire intervenir pour chaque transition le filtrage syntaxique élémentaire.

Le processus d'attribution automatique de catégories grammaticales potentielles, intégré au processus d'enrichissement automatique du vocabulaire (phase initiale), est appliqué à ce moment là :

- soit pour une forme inconnue (c'est-à-dire n'ayant pas de solution grammaticale), de manière à ce que le filtrage syntaxique puisse réduire le nombre de possibilités hors contexte,
- soit dans le cas d'une succession de deux formes incohérentes dont les ensembles solutions ne sont pas vides (une forme incohérente gauche suivie d'une forme incohérente droite), afin d'établir un chemin entre ces deux formes.

Le processus d'enrichissement automatique du vocabulaire peut intervenir à différents stades, dès que l'on a la possibilité d'attribuer de manière non ambiguë une solution morphologique à une forme inconnue. Les nouvelles formes ainsi consignées dans le dictionnaire d'analyse pourront être reconnues lors de l'analyse morphologique, pour les portions de texte suivantes.

Les traitements syntaxiques effectués au cours de l'analyse sont de deux types et correspondent à deux modules :

- le premier, constitué du filtrage positionnel et grammatical, est un traitement portant à l'intérieur d'une transition entre deux formes

successives du texte, sur un arc reliant deux solutions morphologiques correspondantes.

- le second traitement consiste en l'application de schémas de résolution d'ambiguïté; il permet d'une part de tenir compte d'un contexte plus large pour l'expression de contraintes syntaxiques (pouvant porter sur plus de deux formes successives), et d'autre part de disposer d'un formalisme d'expression plus souple pour décrire ces contraintes (plusieurs solutions d'une même forme peuvent être concernées).

Le résultat obtenu après la morphologie est soumis au filtrage syntaxique qui permet d'éliminer un certain nombre d'arcs de la transition définie entre deux formes consécutives. On rappelle qu'un arc relie deux solutions morphologiques des deux ensembles solutions associés à deux formes consécutives. Ce filtrage s'effectue à partir de la consultation d'une matrice de précédence multivaluée qui contient les informations relatives aux relations positionnelles et aux contraintes grammaticales prises en compte. Cette matrice permet donc de valider ou d'infirmer la succession possible de deux solutions morphologiques de deux formes consécutives, et donc de décider de la confirmation ou de l'élimination de l'arc potentiel les reliant.

A l'issue de ce traitement, la résolution des ambiguïtés cataloguées est effectuée, afin de produire un réseau partiellement linéaire, traitable par le processus d'extraction des groupes conceptuels. En effet l'importance de la combinatoire résultante du filtrage syntaxique rend nécessaire une simplification supplémentaire avant toute tentative de traitement de ce réseau, comme par exemple une extraction des groupes nominaux.

Cette simplification s'effectue à ce stade de l'analyse, par application de schéma de résolution d'ambiguïté. Ces schémas sont consignés dans un catalogue défini manuellement, et recensent un certain nombre de cas de figure d'ambiguïté que nous jugeons intéressants pour la linéarisation du réseau résultat. Chaque schéma de résolution se compose de deux parties :

-1) une partie gauche dont la reconnaissance dans le réseau d'entrée (test d'inclusion) est la condition de l'application du schéma de résolution.
Cette partie gauche est, dans le formalisme défini au chapitre précédent, un schéma ambigu (cf. IV.3.6.2.2.).

-2) une partie droite qui représente la portion de réseau à supprimer lors de l'application de ce schéma de résolution.
Cette portion à supprimer est constituée d'un certain nombre d'arcs de la partie gauche, pouvant appartenir à des transitions différentes. Elle constitue un schéma de réseau (cf. IV.3.6.2.1.).

La phrase ci-dessous est tirée d'un corpus technique décrivant les normes d'exploitation et de fonctionnement du CNET, ayant servi de corpus d'expérimentation pour les modules d'indexation automatique et d'interrogation développés dans le cadre du projet IOTA (cf. I.3) :

" ***Les signaux de ligne utilisés devront être identiques à ceux du code de signalisation pour l'accès à l'interurbain manuel, définis à la page 68.*** "

C'est cette phrase qui constituera l'entrée de l'analyseur et qui va nous servir à illustrer notre démarche tout au long de ce chapitre.

Il est à noter que les résultats de cet exemple ont été obtenus à partir d'un dictionnaire d'analyse incomplet, c'est-à-dire qui ne comporte pas l'intégralité du vocabulaire propre à ce corpus. Nous sommes donc bien dans le cas d'une appréhension d'un domaine ouvert.

3. L'analyse morphologique

L'analyse morphologique constitue le premier module de notre analyseur. Elle s'applique directement sur le texte d'entrée sans prétraitement de ce dernier, contrairement à certains systèmes analogues étudiés au chapitre II.

Cette section ne comporte pas la description des fonctions de traitement des formes inconnues, et d'enrichissement automatique du vocabulaire, qui bien qu'étroitement liées à la morphologie, seront présentées dans ce chapitre en tant que modules indépendants (cf. 6. et 7.).

3.1. Définition

L'analyse morphologique (cf. IV.3.2.) effectue une segmentation du texte d'entrée en formes. Les différentes décompositions possibles d'une forme **F**, en racine et désinence, permettent la détermination d'un ensemble solution **S**, qui recense les interprétations linguistiques hors contexte, dans le modèle utilisé. Une interprétation correspond à une solution morphologique **sm** composée

d'un représentant d'unité lexicale, d'une catégorie grammaticale, et d'un ensemble de valeurs grammaticales.

Pour ce faire, on dispose d'un ensemble de données consignées dans un dictionnaire de racines, et de désinences organisées en tables :

- à chaque racine est directement associé un ensemble de renseignements linguistiques qui sont :
  - une catégorie grammaticale,
  - un modèle morphologique,
  - le représentant de l'unité lexicale;

- à chaque désinence (notée "des" sur la figure 2) est directement associé un ensemble de valeurs grammaticales (noté "act").

La décomposition d'une forme, constituée d'une suite de caractères contigus, s'effectue tout d'abord par la reconnaissance des racines potentielles dans le dictionnaire, puis par l'identification de la désinence correspondante parmi l'ensemble des désinences possibles. Ces désinences sont déterminées par le modèle morphologique associé à la racine reconnue.

La décomposition multiple d'une forme donnée (plusieurs couples racine-désinence possibles) génère une ambiguïté.

La rencontre d'une forme inconnue active le processus de catégorisation automatique qui permet de proposer un ensemble solution en fonction de la morphologie de la forme, sans provoquer un arrêt de l'analyse. Nous détaillons ce traitement plus avant dans ce chapitre (cf. V.6.5).

Nous pouvons représenter schématiquement la décomposition d'une forme par la figure ci-après :

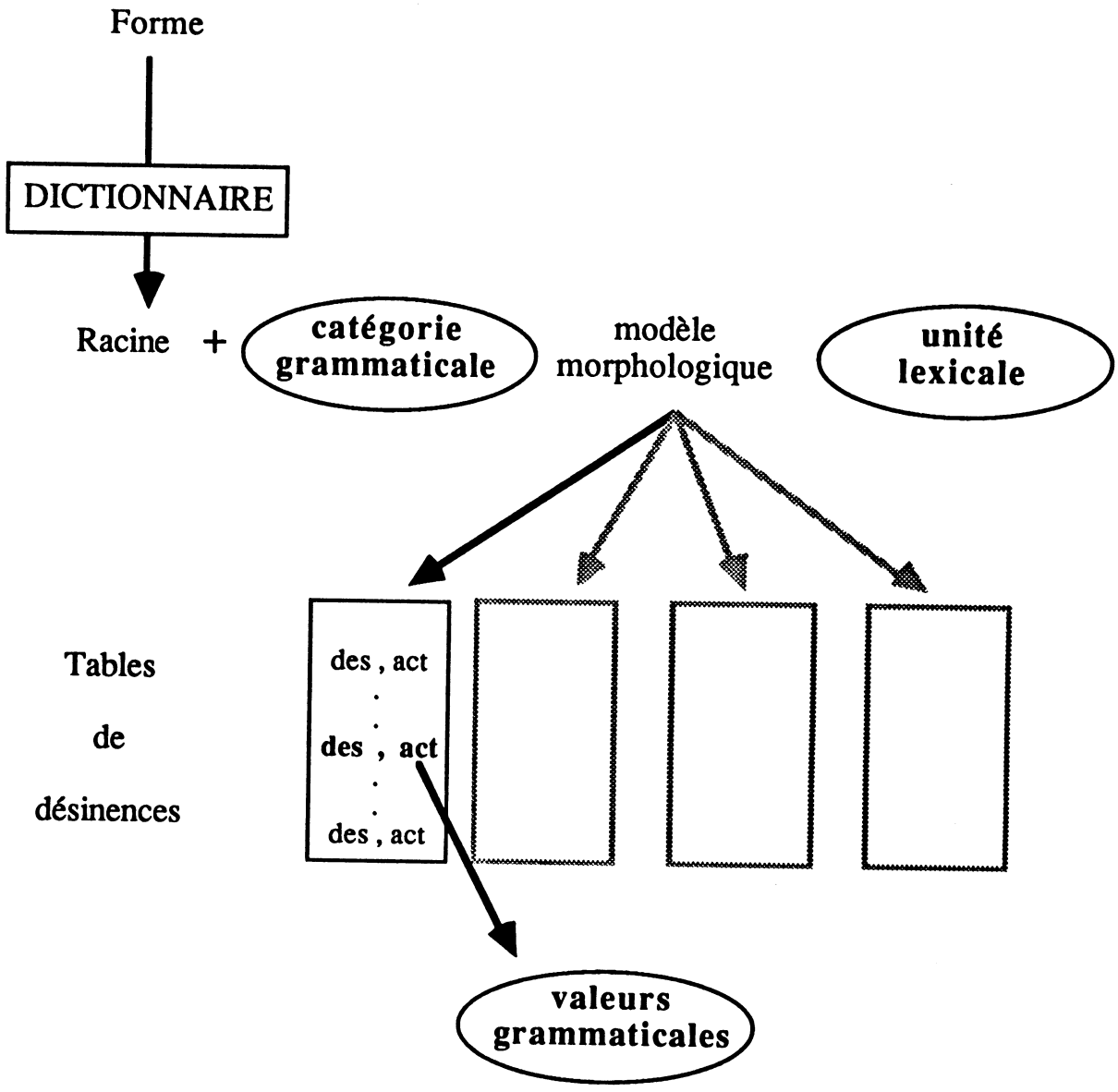


Figure 2 : *Décomposition d'une forme*

3.2. Caractéristiques

Cette analyse morphologique comporte les caractéristiques suivantes qui conditionnent ses performances qualitatives et son efficacité :

- progression caractère par caractère sans retour arrière,
- prise en compte directe des mots composés,

- traitements particuliers lors de la lecture de la chaîne d'entrée permettant la prise en compte des nombres ou des sigles, et le traitement des caractères en majuscule en début de phrase.

3.2.1. Analyse sans "retour arrière"

La progression dans le texte d'entrée sans "retour arrière", caractère par caractère, qui permet de produire en une seule consultation du dictionnaire, l'ensemble des décompositions d'une forme, constitue la principale caractéristique de l'analyseur morphologique.

Cette progression est autorisée par l'organisation du dictionnaire d'analyse en arbre lexicographique (cf. 3.3.2.1.), et par la programmation des tables de désinences. Cette stratégie permet d'éviter l'utilisation de mécanismes de "retour arrière", si fréquents dans ce type d'analyse, et ainsi de minimiser le temps d'exécution de l'analyseur. La figure ci-après permet d'en schématiser le fonctionnement :

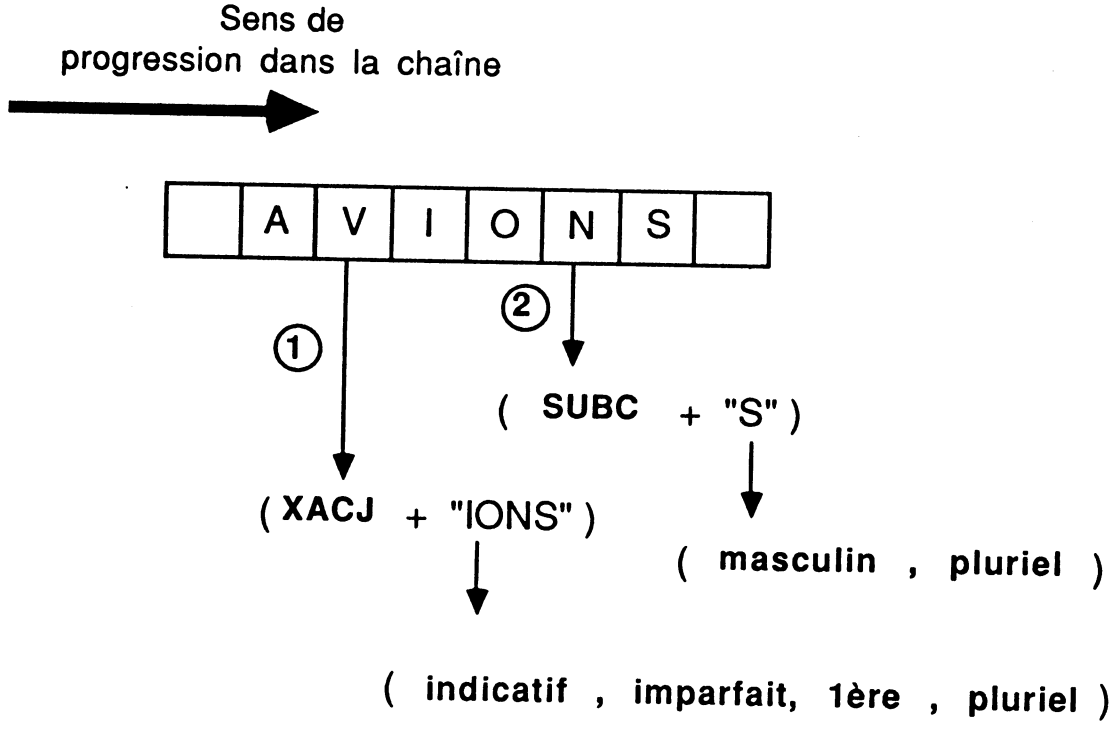


Figure 3 : *Progression de l'analyse*

3.2.2. Reconnaissance des mots composés

La prise en compte des mots-composés, très fréquents en français, est autorisée par la définition paramétrable de deux ensembles de séparateurs de formes (cf. IV.3.1.-e) :

- un ensemble de séparateurs de formes, qualifiés de "non-impératifs" , qui interviennent dans la structuration des mots composés et qui permettent donc leur reconnaissance s'ils sont présents dans le dictionnaire d'analyse. Un séparateur non-impératif fait partie intégrante d'un mot composé. Cela signifie que lorsqu'on le rencontre, on essaie systématiquement de poursuivre la reconnaissance dans le dictionnaire. On ne peut reconnaître un mot composé que s'il est déjà connu du système (contenu dans le dictionnaire).

- un ensemble de séparateurs de formes qualifiés d' "impératifs", et qui ne peuvent pas être contenus dans une forme simple ou composée. La rencontre d'un tel séparateur dans une chaîne d'entrée signifie impérativement la fin de la reconnaissance de la forme qui le précède.

Il est bien évident que cette manière de procéder entraîne un alourdissement du processus d'analyse (tentative systématique de poursuite de reconnaissance lors de la rencontre d'un séparateur non-impératif), mais qui est nécessaire à la prise en compte des mots composés. Leur reconnaissance est indispensable en français. Il s'agit d'un choix que nous préférons aux techniques rejetant cette reconnaissance à un stade ultérieur de l'analyse, et faisant intervenir des informations spécifiques telles que des règles de composition. Ce choix nous permet par ailleurs une gestion plus simple des informations contenues dans le dictionnaire (nous ne stockons pas les informations nécessaires à ce type de règles).

C'est pour cela que nous n'avons pas considéré, dans cette version du système à vocation générale, les caractères spéciaux jouant un rôle particulier dans certaines applications et pouvant être assimilés à des séparateurs non-impératifs pour ces applications :

Exemple :

Les caractères "/" ou "." entrent fréquemment dans la constitution des noms de fichiers informatiques :

"*analyse.pascal* " "*/usr/lib/exemple* "

Il suffit pour pouvoir tenir compte, lors de l'analyse, de ces contextes d'application particuliers, d'utiliser la paramétrabilité des deux ensembles de

séparateurs, et donc de mettre ces caractères dans l'ensemble des non-impératifs.

3.2.3. Traitements particuliers lors de la lecture de la chaîne d'entrée

Des traitements particuliers lors de la lecture de la chaîne d'entrée permettent de prendre en compte des phénomènes marginaux :

- Une procédure spécifique permet lors de cette lecture de considérer les caractères " . " et " , " qui appartiennent à l'ensemble des séparateurs impératifs, comme partie intégrante des nombres dans lesquels ils apparaissent :

  Exemple :

  " 9,5 " " 1.234,56 " " 1.000.000 "

- Une procédure similaire permet de reconnaître les sigles constitués d'une alternance de lettres en majuscule et de points :

  Exemple :

  " A.F.C.E.T. " " C.N.R.S. " " S.N.C.F. "

- Un traitement particulier du premier caractère en majuscule d'une forme débutant une phrase, permet après une conversion en minuscule, la recherche de la racine de la forme dans le dictionnaire d'analyse (ce traitement est un peu plus complexe pour les voyelles que pour les consonnes, car il faut alors considérer en plus, chaque possibilité d'accentuation).

Ces traitements spécifiques nécessitent dans certains systèmes, un processus d'acquisition sophistiqué pouvant être assez lourd (cf. II.3.3. les prétraitements effectués dans le système SYDO).

3.3. Les outils morphologiques

L'analyse morphologique nécessite, outre un modèle linguistique, un certain nombre d'ensembles initiaux d'informations qui correspondent aux données linguistiques, à partir desquels elle peut se réaliser. L'organisation fonctionnelle de ces données conditionne l'efficacité des algorithmes d'analyse. Ces différents ensembles d'informations et les fonctions qui permettent de les exploiter, constituent les outils morphologiques. Pour ce processus nous pouvons discerner :

- d'une part l'ensemble des désinences organisées sous formes de tables, et les modèles morphologiques qui permettent de valider les tables contenant les désinences qui correspondent aux différentes dérivations possibles à partir d'un radical. Ces deux ensembles de taille limitée, sont prédéfinis et figés. Leur exploitation consiste en une consultation permettant une comparaison des terminaisons des formes analysées avec les désinences et une affectation de certaines variables grammaticales à partir des valeurs qui y sont associées.

- d'autre part un ensemble de radicaux consignés dans le dictionnaire d'analyse, auxquels sont associées les informations constituant les renseignements linguistiques (grammaticaux et lexicaux cf. 3.3.2.2.). L'organisation de ce dictionnaire doit privilégier les comparaisons établies lors de l'analyse morphologique entre les formes rencontrées dans les textes analysés et celles connues du système, c'est-à-dire déjà présentes dans le dictionnaire, tout en minimisant sa taille. Néanmoins, les opérations habituelles de mise à jour ne doivent pas être trop coûteuses, particulièrement lorsque l'on souhaite réaliser un enrichissement automatique du vocabulaire.

3.3.1. Les modèles morphologiques

A chaque racine du dictionnaire d'analyse est associé un numéro de modèle morphologique qui correspond en fait, à la validation d'une suite de tables contenant des désinences. Une décomposition sera validée par la reconnaissance au sein d'une de ces tables de la désinence recherchée.

Un modèle morphologique permet donc de regrouper l'ensemble des désinences qui peuvent suivre certaines racines. Ces désinences sont réparties dans des tables ayant chacune leur spécificité (tables propres aux désinences verbales, etc...).

Un ensemble de valeurs grammaticales propres à la désinence reconnue est affecté à la forme analysée (GENRE et NOMBRE pour un substantif, PERSONNE et NOMBRE pour un verbe conjugué, etc...).

Pour les tables contenant des désinences verbales, (les plus nombreuses), on affecte en plus un ensemble de valeurs grammaticales (dans les variables MODE et TEMPS), associées globalement à la table : la reconnaissance d'une désinence de la table entraîne l'affectation pour la forme rencontrée, de l'ensemble de ces valeurs grammaticales.

Un modèle morphologique est associé à une racine. Par conséquent, une unité lexicale qui nécessite plusieurs racines différentes, utilisera plusieurs modèles morphologiques.

Exemple :

Le verbe " *voir* " constituant une unité lexicale, nécessite 4 racines (radicaux) dans notre morphologie qui sont :

" *v , verr , voi , voy* "

à chacune desquelles est associé un modèle morphologique qui permet d'engendrer ses différentes formes :

" *v-is , verr-ai , voi-s , voy-ons* , etc... ".

Chacune des désinences utilisées pour obtenir ces formes, permet d'affecter certaines valeurs grammaticales, concernant pour les désinences verbales les variables grammaticales PERSONNE et NOMBRE, qui leur sont associées :

1ère et 2ème personne, singulier pour la désinence "*is*",
1ère personne, pluriel pour la désinence "*ons*",
etc...

Les tables contenant ces désinences permettent de plus d'affecter les variables grammaticales MODE et TEMPS, toujours à l'aide de valeurs prédéfinies qui sont factorisées au niveau de la table :

indicatif, passé simple pour la désinence "*is*"
(indicatif,présent) et (impératif,présent) pour "*ons*"
etc...

Du fait de la prédéfinition des tables de désinences et des valeurs grammaticales qui leur sont associées, et afin d'améliorer les performances de l'analyseur morphologique, ces tables sont entièrement programmées, et donc compilées en même temps que les algorithmes d'analyse.

3.3.2. Le dictionnaire d'analyse

Le rôle du dictionnaire est classique pour le domaine du traitement automatique de la langue naturelle : permettre de consigner, puis de reconnaître, un ensemble de formes d'un sous-ensemble du vocabulaire de la langue.

Son originalité est constituée par une organisation qui permet d'optimiser les accès, de minimiser la taille et d'effectuer une analyse sans mécanisme de "retour arrière" :

- l'optimisation des accès est obtenue en privilégiant les racines les plus fréquentes,
- la minimisation de la taille résulte d'une factorisation des racines sous forme d'arbre lexicographique,
- enfin, c'est cette factorisation en arbre qui autorise l'analyse caractère par caractère, sans "retour arrière".

3.3.2.1. Structure du dictionnaire

Cette structure d'arbre lexicographique, dont l'étude a fait l'objet d'un rapport de DEA [PALM 81], a principalement consisté en une application de la technique de construction d'arbre de recherche "Median split tree" (MST) proposée par B.A. Sheil [SHEI 78], à une structure classique de factorisation développée par M. Kay [KAY 77].Voici une illustration de la technique (MST) :

Exemple :

soit LL = (a, b, c, d, e, f, g, h, i)
une liste d'éléments donnée dans l'ordre lexicographique, que l'ont doit organiser,
et soit LF = (g, i, b, d, e, a, f, h, c)
la liste contenant le même ensemble d'éléments ordonnée selon les fréquences décroissantes.

La technique MST de construction d'un arbre de recherche consiste en une application récursive de l'algorithme suivant :

- 1) noter le premier élément de LF, et le retrancher des deux listes,

- 2) noter l'élément médian de la liste LL, et constituer deux nouvelles liste ordonnées LL' et LL" avec respectivement les éléments dont le rang est inférieur ou égal, et les éléments dont le rang est supérieur à cet élément médian (les éléments de la première liste vont être rangés dans le sous-arbre gauche, et ceux de la deuxième dans le sous-arbre droit).

L'arbre obtenu à partir de ces éléments avec la technique MST est alors :

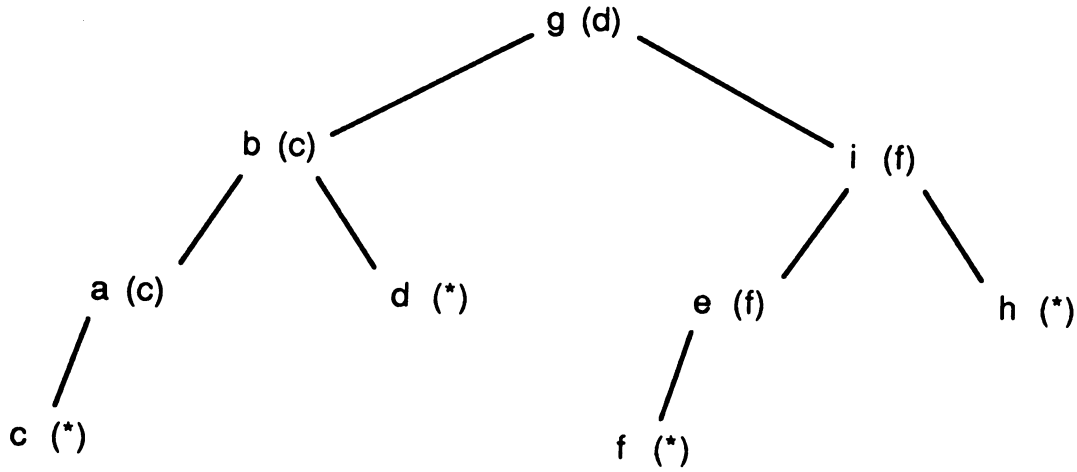


Figure 4 : *Arbre de recherche MST*

Un noeud de cet arbre comporte deux valeurs, une appelée valeur de noeud, représentant l'élément rangé à la place de ce noeud, et l'autre appelée valeur médiane, qui est notée sur cet exemple entre parenthèses, permettant de poursuivre la recherche comme dans les arbres lexicographiques classiques. L'étoile "*" représente l'absence de valeur médiane, et donc l'absence de sous-arbre.

L'algorithme de recherche qui découle de cette organisation est alors le suivant :

SI *l'élément recherché* $\neq$ *la valeur de noeud*
ALORS SI *son rang lexicographique* $\leq$ *au rang de la valeur médiane*
ALORS Poursuivre la recherche dans le sous-arbre gauche
SINON Poursuivre la recherche dans le sous-arbre droit.

C'est cette technique d'arbre de recherche MST qui est directement appliquée pour la construction du dictionnaire, qui se présente donc comme un arbre de racines factorisées (arbre lexicographique) et classées selon un ordre basé sur les fréquences d'apparition. De cette organisation découlent deux conséquences importantes pour l'analyse :

-a) la première déjà énoncée est que la factorisation des racines autorise une analyse progressant caractère par caractère, sans retour-arrière. En effet, on peut explorer toutes les décompositions possibles d'une forme en une seule consultation.

-b) la deuxième est que l'algorithme de recherche d'un élément dans le dictionnaire est équivalent à un parcours d'arbre binaire. En effet, le dictionnaire se présente comme un arbre binaire organisé suivant deux relations d'ordre : un ordre de classement fréquenciel, et un ordre de recherche lexicographique.

De plus, cette structure permet une gestion simple du dictionnaire, principalement pour la fonction d'ajout, puisque dans la technique MST l'ajout d'un nouvel élément (dont la fréquence d'apparition est donc minimale) se fait en feuille.

3.3.2.2. Contenu du dictionnaire

Le contenu du dictionnaire est constitué d'une part, par l'arbre lexicographique factorisant les racines, et d'autre part, par les informations linguistiques associées à chacun de ces radicaux. Ces informations linguistiques sont de deux types :

- grammatical :
  à chaque racine sont associés une catégorie grammaticale et un modèle morphologique permettant de retrouver les dérivations possibles à partir de cette racine,

- lexical :
  à chaque racine est associée une référence à un représentant d'unité lexicale permettant de retrouver le représentant lexical qui est la forme normalisée :
  la forme au masculin singulier pour les substantifs,
  la forme à l'infinitif pour les verbes,
  etc...

La modélisation du contenu du dictionnaire dans le formalisme Z_O défini dans [ABRI 74], que nous donnons ci-après, permet de définir les associations entre ces différents ensembles d'informations :

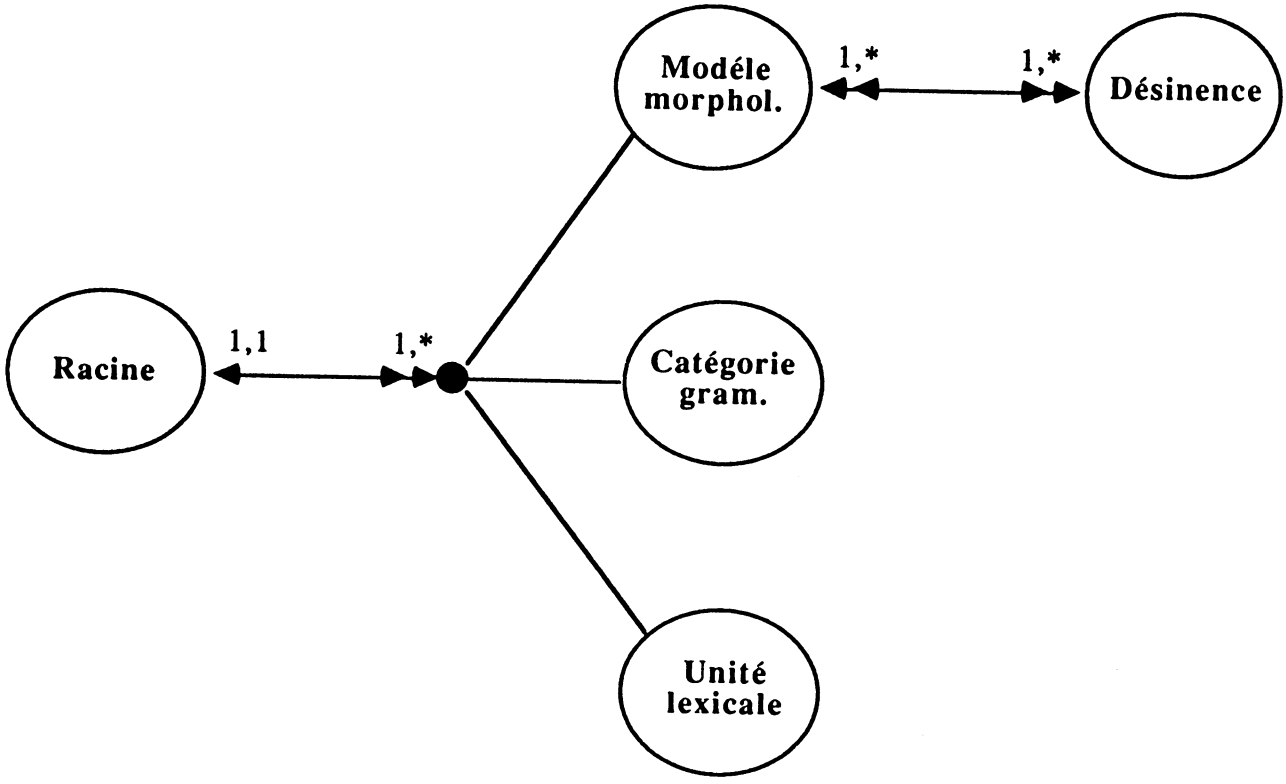


Figure 5 : *Modélisation Zo du contenu du dictionnaire*

les notations des relations d'association sont :

1 , * pour une association multivaluée totale (flèche double),

1 , 1 pour une association monovaluée totale (flèche simple),

et • représente le produit cartésien.

3.3.2.3. Initialisation du dictionnaire

Contrairement aux modèles morphologiques et aux tables de désinences, les données linguistiques consignées dans le dictionnaire d'analyse ne sont ni figées ni aisément circonscriptibles. Leur acquisition constitue le principal problème lié à la constitution du dictionnaire, et en particulier la détermination pour ces données d'un noyau minimum permettant une utilisation satisfaisante

de ce dictionnaire; étant entendu qu'il est toujours enrichissable (manuellement ou automatiquement).

Nous verrons lors de la présentation du processus de catégorisation automatique (cf. V.6) que les données initiales déterminées en fonction des besoins de ce processus peuvent être utilement complétées pour le traitement d'une application particulière en réalisant un apprentissage complémentaire du vocabulaire à partir de textes représentatifs de l'application concernée. Ce complément permet essentiellement de consigner dans le dictionnaire le vocabulaire spécifique de l'application, dont la fréquence d'apparition est significative.

3.4. Analyse morphologique de la phrase exemple

Le résultat de la morphologie pour notre phrase d'entrée est donné sur la figure 6. Les formes analysées sont numérotées, et suivies de leurs solutions morphologiques. Une solution morphologique est en pratique constituée pour les formes reconnues d'un numéro d'unité lexicale, d'une catégorie grammaticale, et d'un ensemble de valeurs grammaticales. Ces différents renseignements linguistiques sont obtenus directement à partir d'informations associées aux racines et aux désinences. La production pour une même forme de plusieurs solutions morphologiques engendre une ambiguïté.

Nous pouvons remarquer à ce stade de l'analyse que plusieurs formes reconnues sont ambiguës. Ces ambiguïtés ne sont pas toutes de même type, et nous pouvons les ranger en deux catégories :

- celles qui diffèrent par la catégorie grammaticale (cas des formes n° 1, 7, 16, 19, 21 et 24),
- celles (cas de la forme n° 25) qui diffèrent par certaines valeurs des variables grammaticales associées.

Les formes n° 12 et 22 relèvent de ces deux cas.

Dans le premier type d'ambiguïté, nous sommes en présence de formes homographes. Dans le deuxième nous avons, soit simplement plusieurs possibilités d'instanciation des variables grammaticales, soit des numéros d'unité lexicale différents, ce qui révèle alors des formes polysèmes.

Pour notre application, nous chercherons à résoudre surtout le premier cas d'ambiguïté, c'est-à-dire les homographies.

1	les	(612	ARTD	MAS,FEM PLU)
		(628	PRPC	MAS,FEM PLU)
2	signaux	(2840	SUBC	MAS PLU)
3	de	(689	PREP)	
4	ligne	(2798	SUBC	FEM SIN)
5	utilisés	(2683	VBPA	MAS PLU)
6	devront	(114	VBCJ	IND FUT 3P PLU)
7	être	(11	XEIF)	
		(2026	SUBC	MAS SIN)
8	identiques	(2851	ADJQ	MAS,FEM SIN)
9	à	(685	PREP)	
10	ceux	(641	PRDM	MAS PLU)
11	du	(1972	ARTC	MAS SIN)
12	code	(2743	VBCJ	SUB PRE 1P,3P SIN)
		(2743	VBCJ	IND PRE 1P,3P SIN)
		(2743	VBCJ	IPF PRE 2P SIN)
		(3260	SUBC	MAS SIN)
13	de	(689	PREP)	
14	signalisation	(	mot inconnu)	
15	pour	(697	PREP)	
16	l'	(612	ARTD	MAS,FEM SIN)
		(628	PRPV	MAS,FEM SIN)
17	accès	(	mot inconnu)	
18	à	(685	PREP)	
19	l'	(612	ARTD	MAS,FEM SIN)
		(628	PRPV	MAS,FEM SIN)
20	interurbain	(	mot inconnu)	
21	manuel	(3018	SUBC	MAS SIN)
		(3019	ADJQ	MAS SIN)
	,	(	VIRG)	
22	définis	(2887	VBCJ	IND PRE 1P,2P SIN)
		(2887	VBCJ	IND PSS 1P,2P SIN)
		(2887	VBCJ	IPF PRE 2P SIN)
		(3260	VBPA	MAS PLU)
23	à	(685	PREP)	
24	la	(612	ARTD	FEM SIN)
		(628	PRPVFEM	SIN)
		(2104	SUBC	MAS SIN,PLU)
25	page	(2784	SUBC	FEM SIN)
		(2785	SUBC	MAS SIN)
26	68	(0	NOMB)	
	.	(	PNTF)	

Figure 6 : *Résultat de la morphologie*

La résolution des polysémies n'est intéressante que lorsque les numéros d'unité lexicale sont différents, c.a.d. lorsque l'on est en présence de plusieurs concepts (cas de la forme n° 25, qui est polysème). Sinon, nous considérerons que nous n'avons pas une véritable ambiguïté, ou plutôt que la résolution de celle-ci n'est pas intéressante dans notre contexte d'application, puisqu'il s'agit d'un même concept (cas des trois premières interprétations de la forme n° 12). Par simplification, nous regrouperons alors dans ce cas, les différentes solutions morphologiques produites en une seule.

D'autres formes non reconnues par l'analyse sont signalées par l'étiquette "mot inconnu". Ce sont les formes dont l'ensemble solution associé est vide : cas des formes n° 14, 17 et 20. Un traitement spécifique est effectué pour ces formes non reconnues (mots inconnus du système) afin de leur attribuer un ensemble d'interprétations potentielles (cf. V.6.5).

A chaque forme du texte d'entrée est donc associé, à ce stade de l'analyse, l'ensemble de ses solutions morphologiques, y compris pour les mots inconnus. Le texte d'entrée est donc entièrement segmenté en formes interprétées.

On rappelle que pour le réseau des solutions morphologiques correspondant, toutes les transitions entre deux formes consécutives sont potentiellement complètes (cf. IV.3.3.2 -d).

La figure 7 permet d'appréhender la combinatoire du réseau potentiel des solutions morphologiques. Nous n'y avons noté que les catégories grammaticales, afin d'en faciliter la lecture.

C'est ce réseau dont les sommets sont les solutions morphologiques déterminées par l'analyse, qui constitue l'entrée du processus de filtrage syntaxique.

3.5. Limites de l'analyse morphologique

Les principales limitations de notre analyse morphologique sont liées à son principe de fonctionnement : progression caractère par caractère sans retour arrière et production en une seule consultation du dictionnaire de toutes les décompositions potentielles d'une forme.

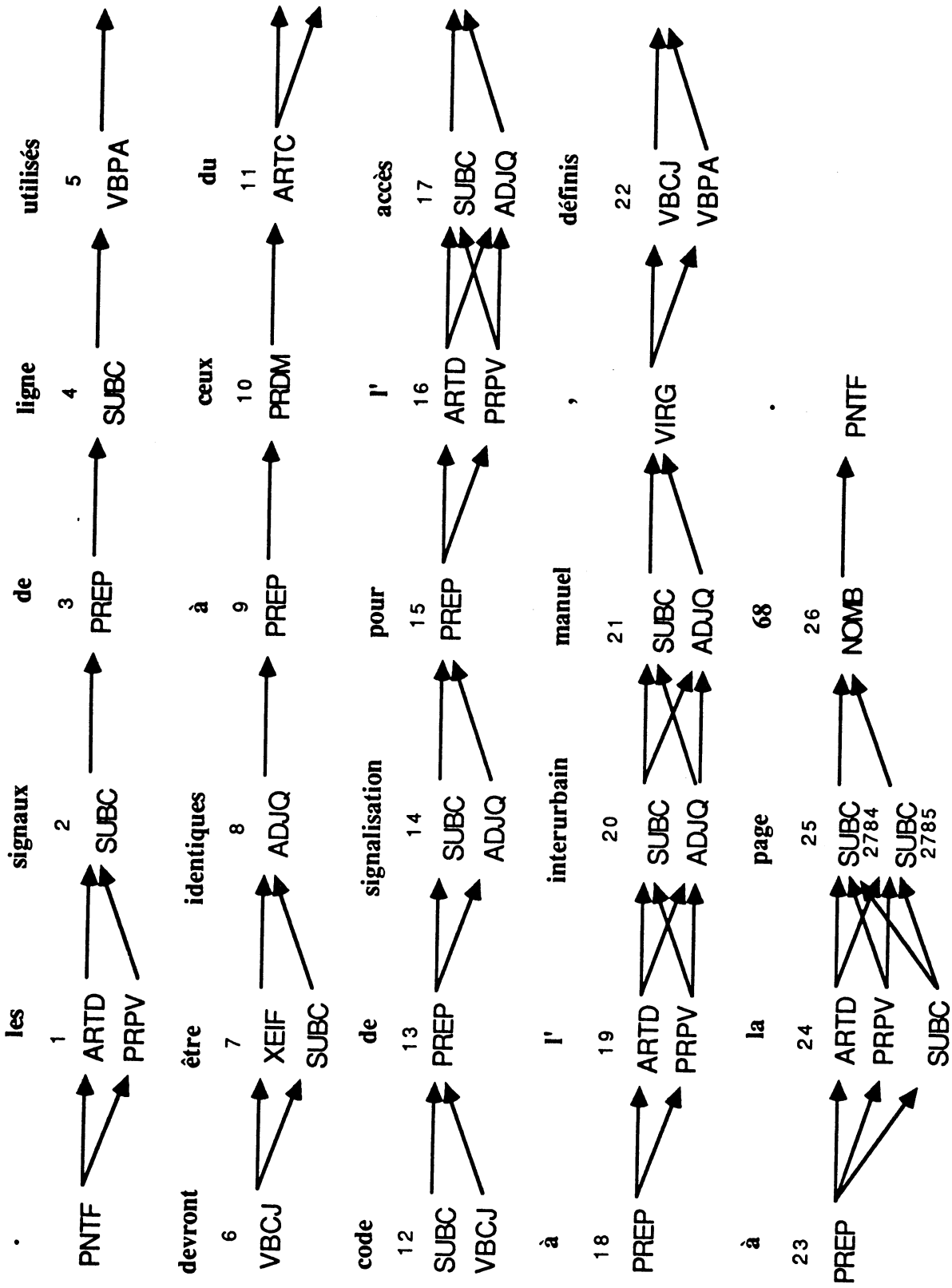


Figure 7 : *Réseau potentiellement complet*

Tout d'abord, nous n'effectuons au cours de l'analyse morphologique, aucune reconnaissance particulière des préfixes. Cette non reconnaissance spécifique des préfixes nous enlève la possibilité de les interpréter directement (privation, répétition, etc...), mais simplifie le schéma de décomposition d'une forme (cette non reconnaissance n'est pas gênante dans notre type d'application).

Plus généralement, dans cette analyse il n'y a pas de traitement spécifique pour les affixes, qui ne sont pas reconnus en tant que tels :

- les préfixes sont considérés comme faisant partie intégrante des racines,
- les suffixes sont tantôt compris dans les racines, tantôt dans les désinences.

Ensuite, lors de la rencontre d'un mot composé, nous n'effectuons que la reconnaissance du mot composé, sans proposer l'interprétation constituée des composants isolés. Si cette limitation n'est pas gênante pour les éléments des classes fermées constituant l'essentiel des données initiales du dictionnaire (nous avons même utilisé cette particularité pour "créer" de nouveaux articles qualifiés de composés, cf. IV.2.2), elle impose une certaine discipline lors de l'ajout de mots composés et plus encore pour les locutions (pluriel des mots composés, locutions divisibles en mots en fonction du contexte).

4. Le filtrage syntaxique

A l'issue de l'analyse morphologique et éventuellement du processus de catégorisation automatique (cf. V.6), chaque forme analysée possède un ensemble solution, noté S, regroupant ses différentes interprétations possibles. La définition de transitions entre ces formes permet de constituer un premier réseau résultat, qui est le réseau morphologique (cf. exemple figure 7). A ce stade de l'analyse, chaque transition est complète, c'est-à-dire que tous les arcs potentiels entre deux solutions morphologiques de deux formes consécutives sont définis, ce qui provoque l'explosion combinatoire des interprétations possibles pour une portion de texte analysée (cf. IV.3.5. propriété 3).

L'objet du filtrage syntaxique est la réduction, à faible coût, d'une partie de cette combinatoire. Son principe est basé sur l'exploitation des relations positionnelles de la langue (possibilités de succession de deux catégories grammaticales), et de certaines contraintes grammaticales "proches" (accords grammaticaux de deux formes consécutives).

4.1. Les relations positionnelles

Pour la définition du modèle linguistique nous avons utilisé une classification des mots de la langue que nous avons appelée catégorisation grammaticale (cf. IV.2.2.). Cette classification a pour but de traduire un comportement syntaxique de ces mots, c'est-à-dire les différentes possibilités d'agencement des uns par rapport aux autres. Ce sont les possibilités d'agencement des catégories grammaticales au sein de phrases qui constituent les relations positionnelles de la langue. Elles ne nécessitent pas l'interprétation sémantique de ces phrases.

A l'occasion de la définition du modèle linguistique, nous avons effectué un certain nombre de choix consistant à diversifier les catégories grammaticales des classes fermées, ou à faire certains regroupements en créant de nouvelles classes. Ces choix ont été dictés par le rôle syntaxique joué par les éléments de ces classes ainsi que par les objectifs que nous nous sommes fixés,

et en particulier une résolution simple (peu coûteuse) des ambiguïtés grammaticales.

Un mécanisme participant à cette résolution va être l'exploitation des relations positionnelles binaires de la langue, c'est-à-dire les possibilités et surtout les impossibilités, de succession de deux catégories grammaticales :

- deux catégories grammaticales peuvent se succéder lorsqu'un élément de la première peut précéder un élément de la seconde.
- deux catégories grammaticales ne peuvent pas se succéder lorsqu'il est impossible de trouver un élément de la première pouvant précéder un élément de la seconde.

Exemple :

un article défini peut précéder un nom commun, par contre, il ne peut précéder un verbe conjugué.

Ces relations positionnelles sont fortement dépendantes de la classification grammaticale utilisée puisqu'elles s'expriment en terme de possibilité de succession de deux catégories grammaticales. Pour établir ces relations nous devons donc considérer tous les couples possibles constitués de deux catégories grammaticales.

Cette binarité, possibilité ou impossibilité de succession de deux catégories grammaticales données, ne permet pas dans un nombre de cas, certes petit mais fréquent, de prendre en compte une diversité liée aux accords grammaticaux de la langue. C'est pour moduler cette possibilité de succession que nous avons introduit les contraintes grammaticales.

4.2. Les contraintes grammaticales

Nous entendons par contraintes grammaticales, les restrictions permettant de conditionner les successions de deux solutions morphologiques par la vérification d'un accord en genre, en nombre, et / ou en personne, portant sur les valeurs des variables grammaticales associées aux catégories de ces solutions morphologiques.

Exemple :

Nous avons vu précédemment qu'un article défini peut précéder un nom commun. Cette possibilité de succession sera sujette à la

vérification d'un accord en genre et en nombre entre les deux solutions.

Les contraintes grammaticales permettent donc d'affiner certaines relations positionnelles entre catégories.

Ce type de contraintes, bien que peu nombreuses lorsque l'on ne considère que deux formes consécutives, nous permettent, par une vérification très simple, de déboucher sur une simplification intéressante de la combinatoire de certaines ambiguïtés. Nous avons retenu trois types de contraintes grammaticales qui représentent les règles d'accord suivantes :

-1) accord en nombre et en genre

-2) accord en nombre et en personne

-3) nécessité de la 3ème personne d'un verbe conjugué précédé d'un nom ou d'un adjectif. Nous ne pouvons pas vérifier dans ce cas un accord en nombre sans déterminer le sujet.

Ce sont ces contraintes que nous avons codées avec les possibilités et impossibilités de succession au sein d'une matrice de précédence. Cette technique peut permettre facilement d'étendre ces contraintes en fonction de renseignements disponibles dans d'autres types d'analyses (comme les traits sémantiques).

4.3. La matrice de précédence binaire multivaluée

La matrice de précédence constitue un outil commode, et d'exploitation très simple, pour représenter les relations positionnelles et certaines contraintes grammaticales proches, de la langue française.

Sans remonter à l'utilisation des matrices de précédence dans les techniques de compilation des langages "artificiels" [COLM 70], on peut dire que ces matrices ont été définies, dans le domaine du traitement automatique des langues naturelles, à partir des règles positionnelles du français dans les travaux de l'équipe dirigée par A. Andreewsky [ANDR 73], [DEBI 77] et [FLUH 77], que nous avons présentés précédemment (cf. II.3.5.). Partant du fait que certaines langues, et le français en particulier, sont des langues positionnelles, c'est-à-dire que les mots d'une langue ne peuvent pas apparaître

à n'importe quelle place dans une phrase, mais obéissent à des règles positionnelles en fonction de leur classification dans le modèle linguistique choisi (appartenance à des classes syntaxiques prédéfinies), l'idée de représenter ces relations dans une matrice de précédence afin de traduire simplement une grammaire d'une langue naturelle a été à l'origine de ces travaux. Le choix de matrices de précédence fréquentielles permettait de fournir systématiquement une solution à une ambiguïté : la solution la plus probable.

Nous avons choisi au contraire de ne considérer que des informations certaines qui, si elles ne permettent pas dans tous les cas de fournir une solution, ont l'avantage de ne pas produire d'interprétations erronées. Nous nous sommes limités à une matrice binaire, contrairement aux travaux précédemment cités, car le formalisme que nous avons défini avec les schémas de résolution d'ambiguïté (cf. 5.), permet outre la prise en compte d'un contexte plus large tout comme les matrices d'ordre supérieur à deux, de tenir compte de la configuration de l'ambiguïté. Cette prise en compte étant réalisée au sein de ces matrices par le calcul des fréquences. C'est la raison pour laquelle le processus de filtrage ne constitue qu'un élément de notre approche de résolution des ambiguïté grammaticales.

Notre modèle linguistique est basé sur une classification des mots de la langue en une cinquantaine de catégories grammaticales, ce qui correspond à quelques 2500 couples de catégories grammaticales à considérer. Même si ce nombre est petit par rapport à celui qui serait engendré par des systèmes comportant une classification en 150 ou 200 catégories, il n'était pas question dans le cadre de la réalisation de notre prototype de considérer manuellement chacun de ces couples.

Nous avons utilisé tout d'abord, après avoir fixé manuellement quelques possibilités et impossibilités évidentes, un mécanisme d'apprentissage automatique similaire à celui utilisé dans [FLUH 77], permettant à partir d'un texte résolu manuellement de déterminer les successions possibles de deux catégories grammaticales. Par la suite, lors de l'analyse de textes nous avons recensé de manière incrémentielle les différentes successions nouvelles rencontrées. Enfin au bout d'une expérimentation relativement courte, devant l'absence d'apparition de nouvelles successions, nous avons conclu à une impossibilité pour tous les couples de catégories grammaticales non rencontrées. La matrice de précédence comprend 5 types de valeurs qui sont :

-1) impossibilité de succession, codée 0;

Exemple :

un pronom personnel sujet ne peut précéder un substantif

prps - subc = 0

-2) possibilité non conditionnée de succession, codée 1;

Exemple :

un substantif peut précéder un autre substantif

subc - subc = 1

-3) possibilité conditionnée à la vérification d'un accord en nombre et en personne, codée 2;

Exemple :

un pronom personnel sujet peut précéder un verbe conjugué s'il y a accord en nombre et en personne

prps - vbcj = 2

-4) possibilité conditionnée à la vérification d'un accord en genre et en nombre, codée 3;

Exemple :

un adjectif qualificatif peut précéder un substantif s'il y a accord en genre et en nombre

adjq - subc = 3

-5) possibilité conditionnée à la présence de la valeur 3ème personne pour la variable nombre d'un verbe conjugué, codée 4.

Exemple :

un substantif peut précéder un verbe conjugué, mais celui-ci doit être à la 3ème personne

subc - vbcj = 4

La figure 8 représente un extrait du contenu de la matrice de précédence multivaluée dont l'intégralité est donnée en annexe (cf. annexe 4) :

	subc	adjq	prps	vbcj
subc	1	1	1	4
adjq	3	3	1	4
prps	0	0	0	2
vbcj	1	1	1	0

Figure 8 : *Extrait de la matrice de précédence multivaluée*

4.4. Le filtrage positionnel et grammatical

Nous utilisons la matrice de précédence comme un filtrage syntaxique (positionnel et grammatical) permettant de valider les interprétations de deux formes consécutives (une interprétation correspond à une solution morphologique).

Ce mécanisme permet une simplification relativement importante mais partielle, de la combinatoire du réseau grammatical, comme nous le verrons plus avant (cf. figure 9).

Ce processus de filtrage positionnel et grammatical s'applique en même temps que la morphologie, plus exactement dès que deux formes consécutives sont analysées morphologiquement. Nous disposons en effet à cet instant de toutes les informations nécessaires pour réaliser ce filtrage : les solutions morphologiques déterminées pour chaque forme comprennent une catégorie grammaticale et les valeurs des variables grammaticales associées. Ceci nous évite d'une part un stockage inutile de solutions morphologiques "erronées" et d'autre part d'avoir à retrouver ultérieurement des informations disponibles précédemment : nous appliquons le filtrage positionnel dès que les solutions morphologiques de deux formes consécutives sont disponibles.

Ce filtrage consiste à appliquer le traitement correspondant à la valeur lue dans la case de la matrice de précédence, déterminée par les deux catégories grammaticales concernées, et dont le résultat est la validation ou la suppression de l'arc reliant les deux solutions morphologiques.

4.5. Filtrage syntaxique de la phrase exemple

La figure 9 représente l'exécution de cette opération sur notre phrase exemple.

La matrice de précédence, dans cet exemple, nous fournit quatre types de résultat d'évaluation qui valident ou invalident les arcs potentiels entre deux solutions morphologiques. Nous avons représenté sur ce schéma ces différents cas par quatre types de flèche :

- les flèches en pointillé signifient une impossibilité absolue de succession des deux catégories grammaticales concernées, et donc une invalidation de l'arc potentiel entre les deux solutions morphologiques.

Exemple :

Un pronom préverbal (forme n°1) ne peut précéder un substantif (forme n°2).

- les flèches en continu signifient au contraire une possibilité de succession inconditionnelle des deux catégories grammaticales concernées, et donc une validation de l'arc potentiel entre les deux solutions morphologiques.

Exemple :

Un substantif (forme n °2) peut précéder une préposition (forme n° 3).

- les flèches en gras signifient également une possibilité de succession des deux catégories grammaticales concernées, conditionnée par une contrainte de cohérence grammaticale entre les deux solutions morphologiques concernées, qui a été vérifiée. Ce test de cohérence porte sur les valeurs des variables grammaticales associées. En occurrence, accord en genre et en nombre des deux formes. L'arc potentiel entre les deux solutions morphologiques est donc validé.

Exemple :

Un article défini (forme n°16) doit s'accorder avec le substantif qu'il précède (forme n° 17).

- la flèche en hachuré signifie que la possibilité de succession entre les deux catégories grammaticales existe, mais que cette possibilité est conditionnée par une contrainte de cohérence grammaticale qui n'a pas été satisfaite dans ce cas. En occurrence, accord en genre et en nombre des deux formes. En conséquence, l'arc potentiel a été invalidé (retiré de la transition concernée).

Exemple :

L'article défini " *la* " (forme n°24) ne peut précéder un substantif masculin (2ème solution de la forme n°25).

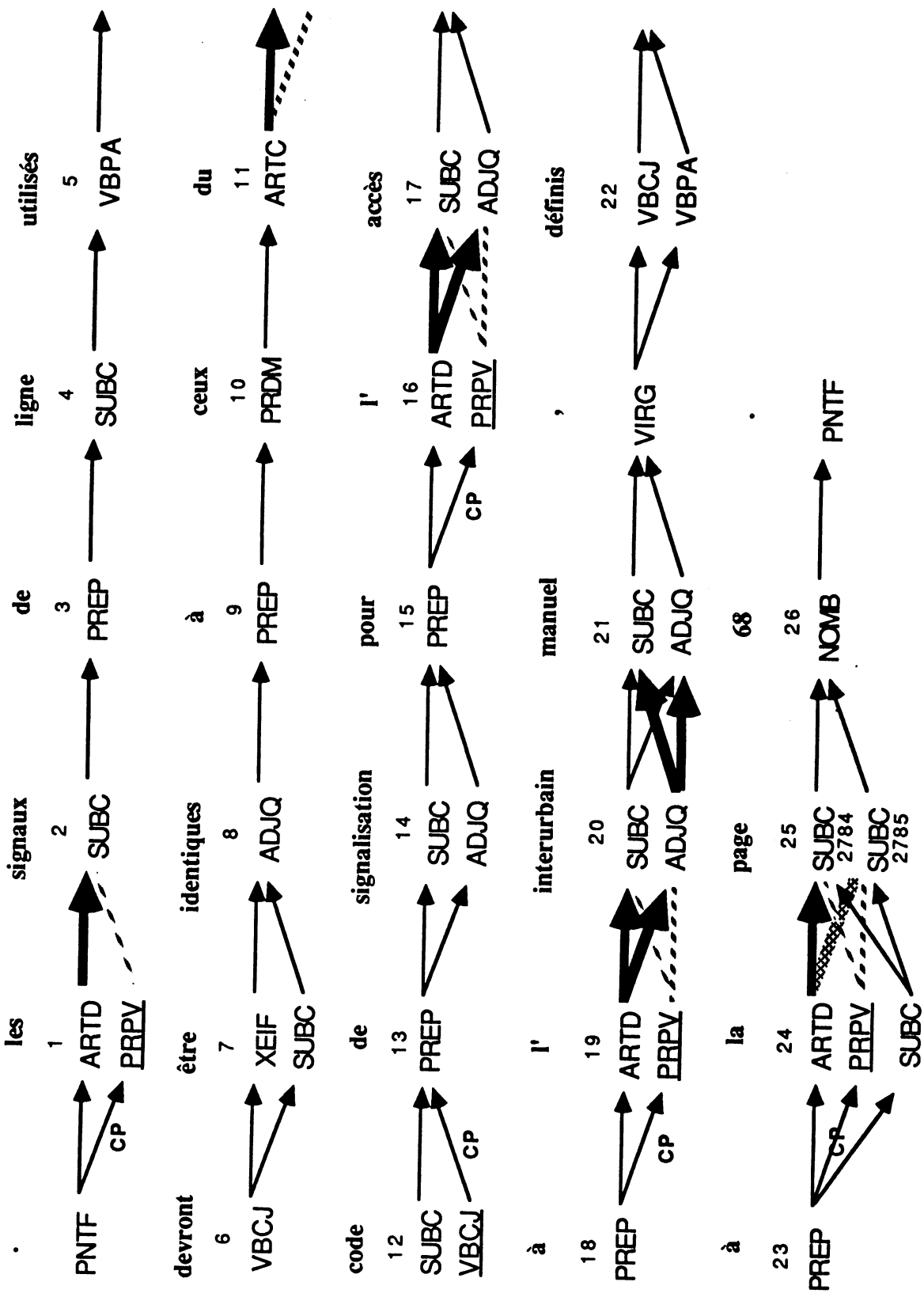


Figure 9 : *Réseau filtré*

On peut remarquer que pour les formes n°12 et 22, les solutions morphologiques ayant une même catégorie grammaticale, et de plus ayant trait au même concept (elles ont le même numéro d'unité lexicale) ont été regroupées. Par contre, en ce qui concerne la forme n°25, les deux solutions ont été conservées puisqu'elles représentent deux concepts différents (les deux numéros d'unité lexicale sont différents).

Nous avons marqué **CP** sur cette trace d'exécution, les arcs constituant des chemins devenus pendants à l'issue de ce traitement. Leur élimination afin d'obtenir un réseau simplifié ne s'effectuera qu'après l'étape suivante qui est l'application des schémas de résolution d'ambiguïté. En effet, les solutions morphologiques isolées (ne figurant plus que sur des chemins pendants), font partie intégrante des ensembles solutions hors-contexte, et peuvent donc concourir à la reconnaissance des schémas de résolution d'ambiguïté (cf. 5.5.1.). Nous avons souligné les solutions morphologiques isolées sur cet exemple.

A l'issue de ce filtrage syntaxique le nombre d'interprétations potentielles de la phrase a diminué. On rappelle que le nombre maximal d'interprétations potentielles est donné par le cardinal du produit cartésien des ensembles solutions (cf. IV.3.4.1 propriété 3). A titre indicatif, ce nombre d'interprétations potentielles à l'issue de la morphologie, correspondant au nombre de chemins entre la première et la dernière forme, s'élevait à 6.144. A l'issue du filtrage syntaxique ce nombre d'interprétations possible a été réduit à 192. En terme d'ambiguïtés, le résultat de la morphologie contenait 12 formes ambiguës. Après filtrage syntaxique, il en reste 8, mais l'une d'entre elles a été réduite (forme n° 25 pour laquelle la solution de catégorie PRPV est isolée). Le nombre de transitions linéaires est de 10 (14 si l'on ne comptabilise pas les chemins pendants), alors qu'il n'était que de 8 à l'issue de la morphologie, sur un total de 28.

4.6. Conclusion sur le filtrage syntaxique

L'originalité de notre approche réside essentiellement dans le fait que nous traitons au même niveau d'analyse des contraintes purement structurelles que sont les relations positionnelles de la langue naturelle et des contraintes de nature grammaticale.

Ce double calcul n'engendre pas de difficultés supplémentaires, puisque nous disposons à ce niveau de toutes les informations nécessaires. De plus la fiabilité des connaissances utilisées (informations contenues dans la matrice) et

la portée limitée (deux formes consécutives) de ce calcul nous autorisent à ne pas revenir sur les résultats obtenus, même si leur relative pauvreté nécessite un processus complémentaire de simplification de la combinatoire du réseau obtenu.

Sur un plan purement pratique, la relative facilité de mise en oeuvre et la simplicité de représentation de cette connaissance syntaxique font de la matrice un outil très commode à manipuler, notamment lors des périodes de mise au point. Un utilitaire de modification de la matrice permet à partir des résultats d'expérimentation de cette dernière, de la faire converger rapidement vers une version suffisamment complète pour être opérationnelle.

Nous avons vu sur l'exemple précédent (cf. figure 9), un résultat de filtrage syntaxique sur un réseau morphologique, et nous avons pu constater qu'un certain nombre d'ambiguïtés grammaticales persistaient après son exécution. La réduction de cette combinatoire résultante est l'objectif du processus de résolution des ambiguïtés grammaticales, qui vient compléter ce filtrage positionnel et grammatical qui ne constitue qu'un débroussaillage du réseau ambigu résultant de la morphologie.

5. L'application des schémas de résolution d'ambiguïté grammaticale

L'objectif du processus d'application des schémas de résolution d'ambiguïté grammaticale est de résoudre certaines des ambiguïtés grammaticales persistantes après le filtrage syntaxique, de manière à obtenir un réseau résultat le plus linéaire possible (cf. IV.4). Ce réseau constitue ensuite l'entrée du processus d'extraction des Groupes Conceptuels.

Rappelons, avant d'aller plus avant, notre choix délibéré de nous restreindre à une analyse de surface de la langue naturelle, sans rechercher les structures syntaxiques complètes des phrases des textes analysés. Ceci afin d'obtenir un outil d'analyse performant, entièrement automatique, capable d'appréhender des univers textuels ouverts et d'être aisément adaptable à un corpus particulier. Dans le contexte de la reconnaissance des groupes nominaux tels que définis dans notre modèle, la désambiguïsation complète des phrases n'est pas absolument nécessaire.

C'est ce choix qui nous a conduit à rechercher une stratégie originale de résolution des ambiguïtés grammaticales, ne nécessitant pas, contrairement aux analyseurs syntaxiques classiques que nous avons présentés précédemment (cf. II.3.), une description complexe sous la forme de grammaires de la langue naturelle. Nous avons renoncé en particulier à utiliser des données linguistiques sophistiquées définies a priori, qu'il serait impossible de déduire automatiquement pour les formes nouvelles rencontrées en cours d'analyse, ou qui nécessiteraient des compétences linguistiques hors de portée de l'utilisateur commun d'un tel système (dans le cas où il désirerait enrichir manuellement le vocabulaire). La relative pauvreté de ce modèle linguistique, ne nous permet donc pas une analyse fine des textes en langue naturelle, mais au contraire lui confère un caractère partiel qui correspond à notre objectif final : la reconnaissance des groupes nominaux.

A l'issue du processus de filtrage syntaxique un certain nombre d'ambiguïtés persistent, essentiellement du fait du contexte très restreint (deux formes consécutives) utilisé lors de la définition des relations positionnelles et des contraintes grammaticales. La résolution de ces ambiguïtés nécessite soit un élargissement du contexte pris en compte, soit l'utilisation d'informations

linguistiques autres que celles déterminées lors de l'analyse morphologique. Nous nous sommes restreint à la détermination d'un ensemble d'informations linguistiques déductibles automatiquement de la morphologie de la forme analysée, ce qui exclut tout trait de type sémantique ou même de type syntaxique tels que par exemple les traits de rection des verbes.

Par conséquent nous nous sommes intéressé aux ambiguïtés grammaticales solubles par un élargissement du contexte d'analyse. La caractérisation de ces ambiguïtés s'établit, outre par la description de l'ambiguïté elle-même, par la donnée de son contexte de résolution.

5.1. Définition

Nous avons défini un modèle permettant de représenter et de classifier ces ambiguïtés grammaticales au sein de leur contexte de résolution. Un schéma de résolution d'ambiguïté grammaticale de longueur n, noté SRA^n, comporte l'identification (au sein d'un contexte déterminé) et la solution ou la simplification d'une ambiguïté donnée, et se compose de deux parties :

- une partie gauche qui représente une ambiguïté grammaticale et son contexte de résolution,
- et une partie droite qui permet la résolution de l'ambiguïté.

Dans le formalisme défini au chapitre précédent (cf. IV.3.6.) :

- la partie gauche d'un schéma de résolution d'ambiguïté de longueur n est un schéma ambigu de même longueur noté SA^n; sa reconnaissance dans le réseau permet l'application de sa partie droite;

- la partie droite est constituée par un schéma de réseau de même longueur n, noté SR^n, constitué de n schémas de transition qui contiennent chacun l'ensemble des arcs à retirer des n transitions correspondantes du réseau analysé pour solutionner l'ambiguïté. Nous avons choisi cette solution plutôt que celle consistant à donner l'ensemble des arcs à conserver, pour une plus grande facilité de mise en oeuvre. Ce choix est explicité dans [BERR 85]. Il est important de préciser que les arcs donnés en partie droite, doivent être définis en partie gauche, c.a.d. que chaque transition de la partie droite doit être incluse dans la transition correspondante de la partie gauche.

Un schéma de résolution d'ambiguïté s'écrit :

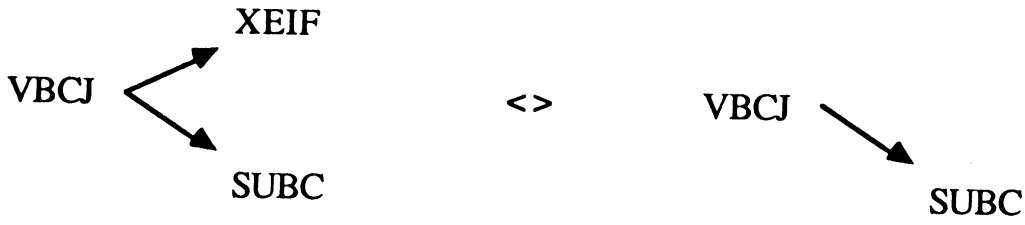
SRA^n :: SA^n <> SR^n

Dans la suite de l'exposé nous utiliserons les notations simplifiées suivantes, qui expriment qu'un schéma S est constitué d'une partie gauche G et d'une partie droite D :

S :: G <> D

Exemple :

Ce schéma de longueur 1, permet de solutionner une ambiguïté concernant l'homographie entre le substantif commun SUBC et l'auxiliaire à l'infinitif XAIF "être", lorsqu'elle est précédée d'un verbe conjugué VBCJ. Cette ambiguïté n'est pas résolue par le filtrage syntaxique.



Les schémas sont définis manuellement et répertoriés au sein d'un catalogue dont la constitution est incrémentielle (cf. 5.4.).

La sémantique d'un schéma de résolution d'ambiguïté grammaticale peut être donnée sous la forme d'une règle de production :

SI **G est reconnue** *ALORS* **appliquer D**

5.2. Applicabilité d'un schéma

La détermination de l'applicabilité d'un schéma de résolution d'ambiguïté s'effectue par une comparaison entre la partie gauche représentant l'ambiguïté au sein de son contexte de résolution et le réseau filtré. La reconnaissance de

cette partie gauche dans une portion de réseau conditionne applicabilité du schéma. Pour caractériser les différents états de cette reconnaissance, nous utilisons les notions d'activation et de validation de ces schémas :

- un schéma est dit "activé" lorsque la reconnaissance de l'inclusion de sa partie gauche dans le réseau est commencée;

- il est validé lorsqu'il est entièrement reconnu.

C'est la validation qui détermine l'applicabilité du schéma de résolution d'ambiguïté sur la portion de réseau concernée.

5.2.1. Activation

La notion d'activation représente l'état d'un schéma en cours de reconnaissance, à une position donnée dans l'analyse du réseau. Cet état, outre l'identification, permet au cours de la progression dans le réseau, de déterminer le stade de la reconnaissance de la partie gauche du schéma activé. Tout schéma du catalogue dont la première transition de la partie gauche est reconnue dans la transition courante du réseau est activé, et est dupliqué dans une liste d'activation. La reconnaissance d'un schéma activé se poursuivra à partir de son occurrence dupliquée dans cette liste.

Plusieurs schémas peuvent être simultanément actifs. On peut avoir également plusieurs occurrences d'un même schéma activées en même temps (à des stades de reconnaissance différents) et donc dupliquées dans la liste d'activation.

5.2.2. Reconnaissance d'un schéma activé

La reconnaissance d'un schéma activé s'effectue transition par transition, en progressant simultanément dans la partie gauche du schéma et dans la portion de réseau considérée. A chaque pas de cette progression on vérifie l'inclusion de la transition courante de la partie gauche dans la transition courante du réseau : c'est-à-dire l'inclusion ensembliste de l'ensemble d'arcs qui constitue le schéma de transition courant de la partie gauche, dans l'ensemble d'arcs constituant la transition courante du réseau (cf. IV.3.6.2.1 propriété 11).

Un schéma est désactivé lorsque la reconnaissance de la transition courante de sa partie gauche a échouée; il est validé lorsque sa partie gauche a été entièrement reconnue dans le réseau.

La figure ci-après permet de visualiser les différentes étapes de la reconnaissance d'un schéma activé, dans une portion du réseau analysé.

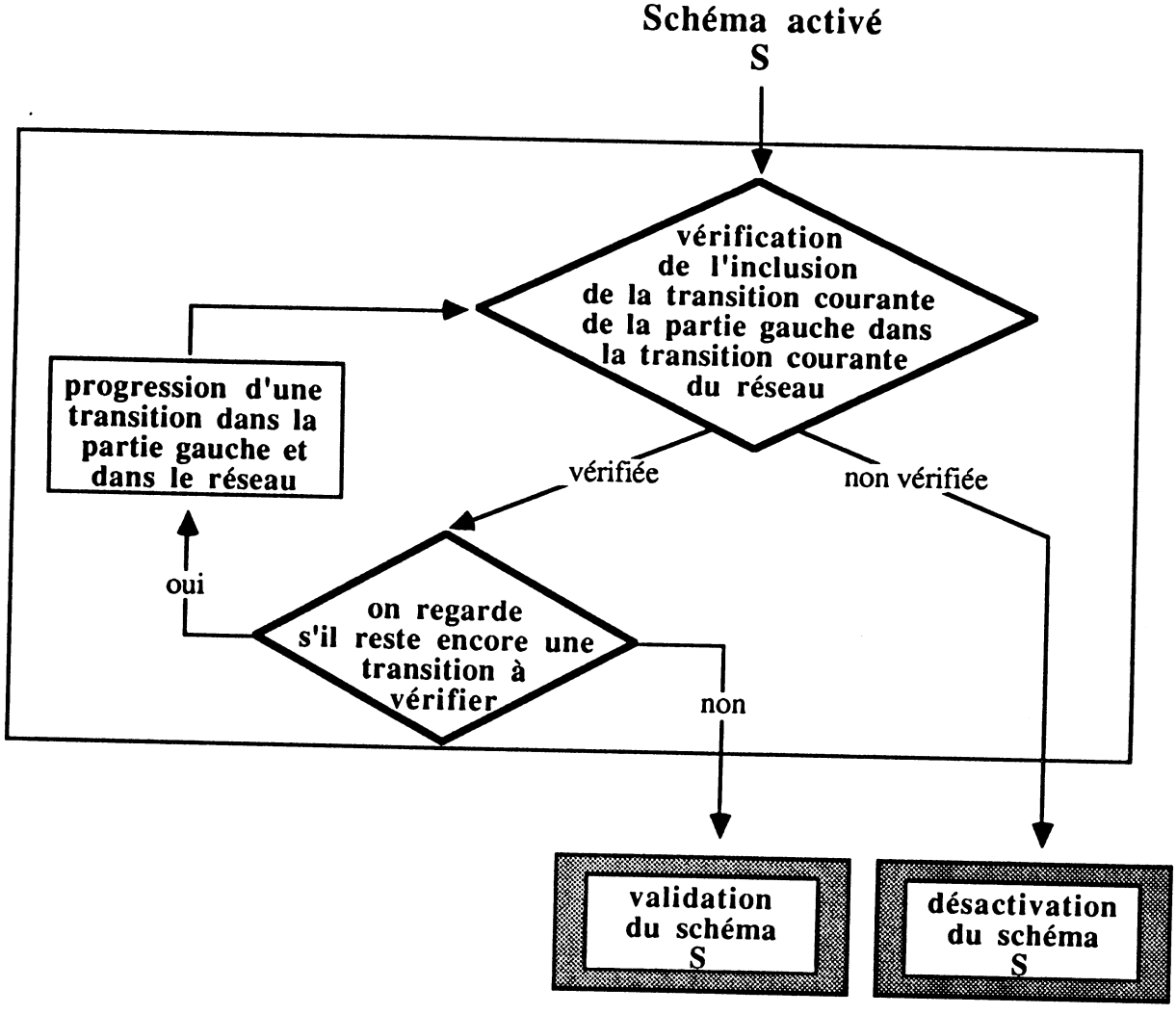


Figure 10 : *procédure de reconnaissance d'un SRA*

Le résultat de cette procédure est soit la validation du schéma si cette reconnaissance a été menée à son terme (jusqu'à la reconnaissance de la dernière transition), soit la désactivation du schéma testé. La désactivation d'un schéma entraîne la suppression de son occurrence de la liste d'activation.

5.2.3. Validation

Pour qu'un schéma de résolution d'ambiguïté de longueur n soit validé, il suffit que l'on ait reconnu l'inclusion de sa partie gauche dans le réseau, c'est-à-dire que les n transitions du schéma ambigu SA^n qui compose sa partie gauche, soient incluses dans n transitions consécutives du réseau (cf. IV.3.6.2.1 propriété 12). Ce qui peut se réécrire de la manière suivante :

Un schéma de partie gauche $SA^n = \{ST_1, ..., ST_n\}$ est reconnu dans un réseau $R_{1,p} = \{T_1, ..., T_i, ..., T_p\}$ si et seulement si :

$$\exists\, i, \text{ tel que } \forall\, k,\ 1 \leq k \leq n \quad \text{on ait } ST_k \subseteq T_{i+k-1} \quad \text{avec } T_{i+k-1} \in R_{1,p}$$

Un schéma validé est applicable sur la portion de réseau analysée, ayant servi à la reconnaissance de sa partie gauche. On parlera pour désigner cette portion de portée de reconnaissance (ou d'application) du schéma sur le réseau.

5.3. Application d'un schéma de résolution d'ambiguïté

La partie droite d'un schéma contient les informations nécessaires à la résolution, ou du moins à la simplification, de l'ambiguïté décrite en partie gauche. L'application d'un schéma de résolution d'ambiguïté réalise donc la simplification décrite en partie droite. Cette simplification consiste en une différence ensembliste entre les transitions de la portion de réseau analysée et les schémas de transition constituant la partie droite du schéma.

L' application d'un schéma de longueur n, dont la partie droite est le schéma de réseau $SR^n = \{ST'_1, ..., ST'_n\}$ sur une portion que l'on peut écrire $R_{i,i+n-1}$, du réseau $R_{1,p}$, se traduit par le retrait dans la portion de réseau $R_{i,i+n-1}$ de tous les arcs définis par la partie droite SR^n.

On peut représenter cette opération par l'expression suivante, dans laquelle le signe ▫ représente l'opérateur de différence ensembliste :

$$R_{i,i+n-1} \text{ ▫ } SR^n$$

que l'on peut traduire par :

pour tout k, $1 \leq k \leq n$ exécuter $T_{i+k-1} \circ ST'_k$.

La figure ci-après permet de visualiser ce mécanisme d'application d'un schéma sur une portion de réseau :

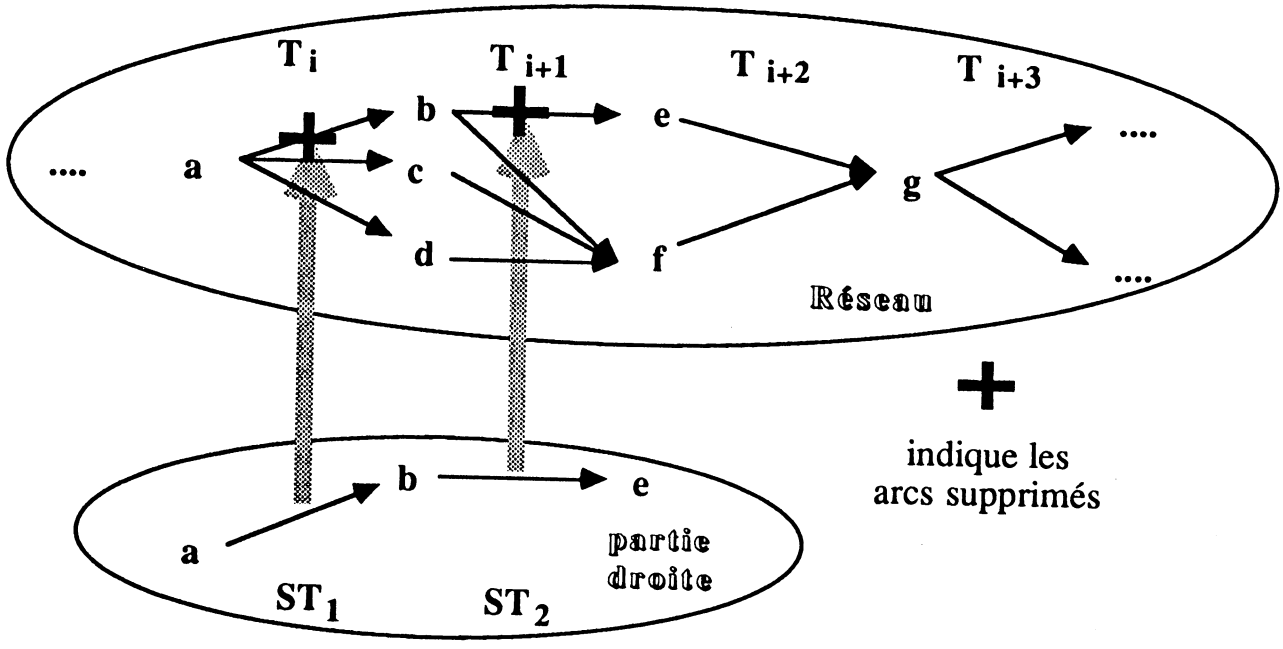


Figure 11 : *Application de la partie droite d'un SRA*

Il est à noter que certains des n schémas de transitions constituant la partie droite d'un schéma de résolution d'ambiguïté grammaticale peuvent être vides : toutes les transitions du réseau qui forment la portée de reconnaissance, et donc d'application d'un schéma, n'ont pas à être simplifiées pour résoudre ou simplifier une ambiguïté. Nous verrons dans la section consacrée à la détermination de la partie droite d'un schéma (cf. V.5.5.2), qu'il est important afin de garantir l'exactitude des résultats obtenus, de donner en partie droite l'ensemble d'arcs minimal permettant de simplifier l'ambiguïté concernée.

5.4. Application des schémas de résolution d'ambiguïté

Nous avons élaboré le processus de résolution des ambiguïtés grammaticales en deux étapes successives. Dans une première approche, nous avons envisagé de retranscrire en partie gauche d'un schéma, l'ambiguïté et le contexte complet de sa rencontre. Cette solution s'est avérée peu satisfaisante vis à vis de nos objectifs d'efficacité liés au choix d'une analyse syntaxique partielle. Nous avons alors élaboré une seconde stratégie permettant une plus grande souplesse de définition d'un schéma de résolution d'ambiguïté, afin de réduire le nombre de schémas nécessaires à une linéarisation satisfaisante du réseau, ainsi que la complexité de ces schémas. Ce résultat s'obtient comme on le verra dans la section suivante (cf. V.5.5.3), au prix d'un effort supplémentaire lors de l'ajout dans le catalogue d'un nouveau schéma. Il est en effet nécessaire afin de toujours garantir l'exactitude des résultats obtenus d'éviter les résolutions contradictoires.

5.4.1. Première approche

Dans une première spécification du processus d'application des schémas de résolution d'ambiguïté, la démarche retenue, développée dans [BERR 85 et 86] et [PALM 85], consiste à tester la concordance totale ("pattern matching"), entre la partie gauche d'un schéma et une portion de réseau, afin de déterminer son applicabilité.

On garantit l'application d'un seul schéma à la fois (le premier validé parmi l'ensemble des schémas activés en même temps) en interdisant tout recouvrement "suffixe" des parties gauches des schémas (même schéma de transition final). Une application a pour conséquence de vider la liste d'activation. Le processus de reconnaissance des schémas repart à partir de la transition suivant la dernière modifiée par l'application du schéma dans le réseau. Les portées d'application des schémas sont disjointes.

Ces contraintes permettent de garantir l'exactitude des résultats obtenus mais limitent fortement l'efficacité du processus du fait le la complexité des parties gauches (devant correspondre exactement aux portions de réseau), ce qui entraîne une multiplication des schémas (chaque variante d'une ambiguïté doit correspondre à une partie gauche différente).

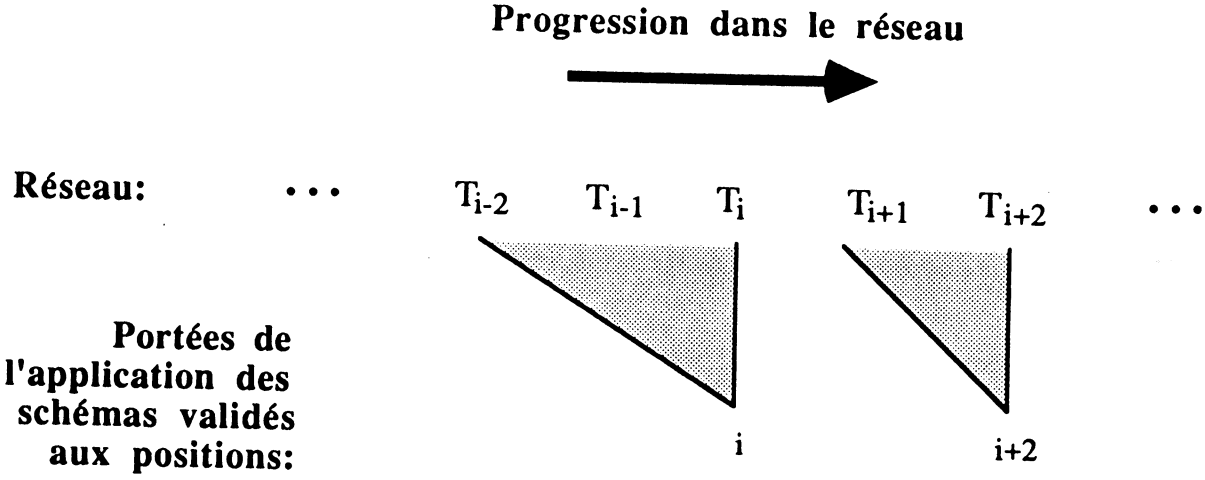


Figure 12 : *portées d'application des SRA*

Ce processus est équivalent à l'application d'un transducteur d'état fini déterministe : les parties gauches des schémas correspondent à la donnée des chaînes de transition permettant de passer d'un état initial à un état final. La construction de l'automate à partir d'un ensemble de schémas est immédiat.

Le bornage des parties gauches par des "points d'ancrage" (formes virtuelles dont l'ensemble solution ne comporte qu'un élément) permet dans cette démarche, de déterminer aisément les parties gauches de schémas, en extrayant du réseau filtré les portions ambiguës. Un mécanisme analogue est utilisé par G. Kallas, dans le cadre du projet SYDO présenté au Chapitre II (cf. II.3.3); la multiplication des configurations est alors fortement limitée par le modèle linguistique qui ne comporte que 11 catégories et par l'approche probabiliste utilisée, [KALL 87].

Le formalisme de ces schémas autorise une détermination aisée de ces ambiguïtés, et par la même, facilite la constitution manuelle du catalogue recensant les schémas de résolution d'ambiguïté. Lors de la création d'un nouveau schéma, il suffit de vérifier le non recouvrement "suffixe" de sa partie gauche avec celles des schémas existants. Cette démarche a donné lieu a un prototype qui a permis de vérifier la faisabilité d'une telle résolution en utilisant un contexte grammatical borné des ambiguïtés. Néanmoins, la multiplication et la complexité excessives des schémas constitués entraîne un alourdissement sensible des performances de l'analyseur, contradictoire avec les objectifs que nous nous sommes fixé avec le choix d'une analyse de surface. Afin de solutionner ces difficultés, nous avons restreint les contraintes portant sur la définition des schémas exposée précédemment, et réexaminé le principe

de la correspondance exacte de la partie gauche dans le réseau tout en maintenant ce déterminisme. C'est ce qui nous a conduit à retenir pour la reconnaissance des parties gauches une correspondance partielle (l'inclusion), et en conséquence à définir une nouvelle stratégie d'application des schémas.

5.4.2. Stratégie d'application des schémas

L'inclusion de la partie gauche d'un schéma dans une portion de réseau, ne permet pas, même lorsque l'ambiguïté décrite dans le schéma est résolue, d'avoir l'assurance de résoudre celle présente dans la portion de réseau analysée. L'application du schéma peut en effet, ne permettre qu'une simplification de l'ambiguïté de cette portion. La résolution complète de cette ambiguïté nécessite alors l'application sur cette même portion de plusieurs schémas décrivant avec leurs parties gauches l'ensemble de sa configuration. Ce recouvrement des portées d'application des schémas de résolution d'ambiguïté permet d'obtenir des résultats équivalents à ceux obtenus avec la première méthode, tout en réduisant considérablement le nombre de schémas du catalogue, ainsi que leur complexité.

Le réseau filtré est parcouru transition par transition de gauche à droite. La reconnaissance des schémas s'effectue au fur et à mesure de cette progression. Plusieurs schémas, ou occurrences d'un même schéma, peuvent donc être activés en même temps et dupliqués dans la liste d'activation. Leurs portées respectives de reconnaissance sur le réseau, se recouvrent alors. De même, plusieurs schémas peuvent être validés au même moment de l'analyse, lorsque la reconnaissance du dernier schéma de transition de leur partie gauche s'effectue sur la même transition du réseau. Leur application s'effectue alors sur des portions qui se chevauchent. Le fait de reconnaître l'inclusion de la partie gauche d'un schéma de résolution d'ambiguïté dans une portion de réseau, afin de déterminer son applicabilité, renforce ce phénomène.

Chaque occurrence de schéma validé est consignée dans une liste de validation et retirée de la liste d'activation. Avant chaque progression dans le réseau, donc après que les tests de reconnaissance sur la transition courante du réseau sont terminés, les schémas figurant dans la liste de validation sont appliqués, et la liste de validation est vidée. Ainsi, cette application ne risque pas de perturber la reconnaissance des schémas activés. La validation simultanée de plusieurs schémas signifie que les conditions d'applicabilité de leurs parties gauches sont compatibles (inclusion du dernier schéma de transition de leur partie gauche dans une même transition du réseau). Il faut donc s'assurer que leur application ne débouchera pas sur des résolutions contradictoires.

La démarche retenue est l'application systématique de tous les schémas validés. L'application d'un schéma ne doit pas provoquer la désactivation de ceux en cours de reconnaissance afin de permettre une application simultanée de plusieurs schémas. Cette stratégie est équivalente à une application en parallèle de tous les schémas applicables pour une position donnée dans le réseau.

Du fait du chevauchement des portées d'application des schémas de résolution d'ambiguïté validés, le retrait d'un arc donné du réseau peut être tenté plusieurs fois. Cela n'est pas gênant dans la mesure où le résultat final est identique quel que soit l'ordre d'application des schémas. La figure 13 ci-après illustre ce phénomène de chevauchement des portées d'application des schémas.

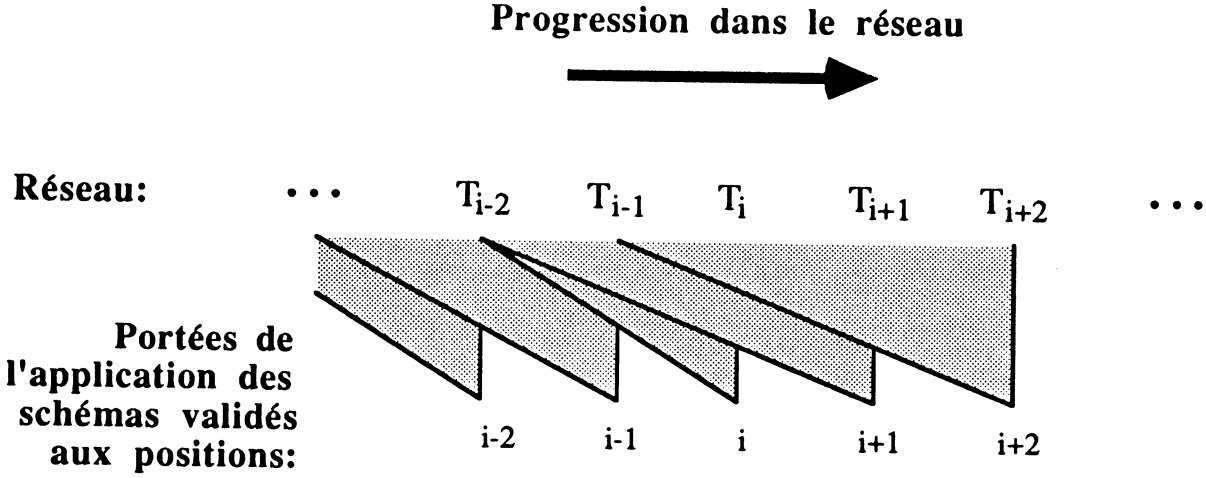


Figure 13 : *chevauchement des portées d'application des SRA*

Cette stratégie d'application peut se résumer ainsi : à chaque étape de la progression dans le réseau analysé, l'application des schémas validés réalise les différences ensemblistes entre les transitions de la portion de réseau concernée, avec les schémas de transition correspondants, qui résultent de l'union des schémas de transition constituant les parties droites des schémas appliqués.

A l'issue de ce processus d'application les chemins pendants et les solutions morphologiques isolées sont éliminés du réseau par un processus de "nettoyage" (cf. 5.8.).

5.5. Catalogage des schémas

Il permet de recenser l'ensemble des schémas définis. Le catalogue initial est le résultat d'un processus d'apprentissage à partir d'un échantillonnage de textes, et de connaissances antérieures constituées par des configurations classiques d'ambiguïtés. En cours d'expérimentation, ce même processus permet de compléter le catalogue au fur et à mesure de l'analyse de nouveaux textes. La figure 14 permet d'en schématiser le fonctionnement :

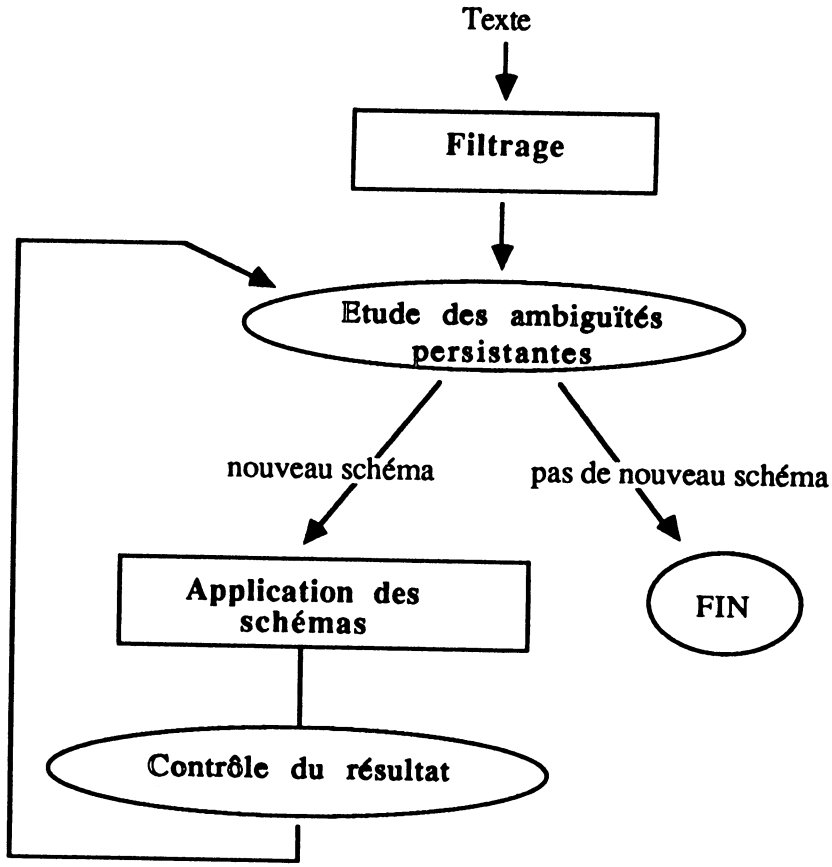


Figure 14 : *Apprentissage des schémas de résolution d'ambiguïté*

Chaque ambiguïté rencontrée est étudiée pour savoir s'il est possible d'en déduire une constante de résolution (ou de simplification) pouvant déboucher sur la création d'un nouveau schéma. Ce processus d'apprentissage est contrôlé par le gestionnaire qui à l'issue de l'application des schémas, vérifie le résultat obtenu et étudie les configurations d'ambiguïtés persistantes, de manière à essayer d'extraire de nouvelles constantes de résolution.

L'ordre de définition de ces schémas n'a pas d'importance puisque la stratégie d'application de ces schémas n'en tient pas compte. La constitution du catalogue s'effectue manuellement de manière incrémentielle.

5.5.1. Détermination de la partie gauche d'un schéma

Lors de la création d'un schéma, la partie gauche doit représenter l'ambiguïté à solutionner et son contexte de résolution (on rappelle que cette ambiguïté peut ne représenter qu'une partie de celle de la portion de réseau considéré). C'est ce contexte qui doit permettre la résolution : une même ambiguïté peut admettre des solutions différentes lorsqu'elle est rencontrée dans des configurations différentes.

Exemple :

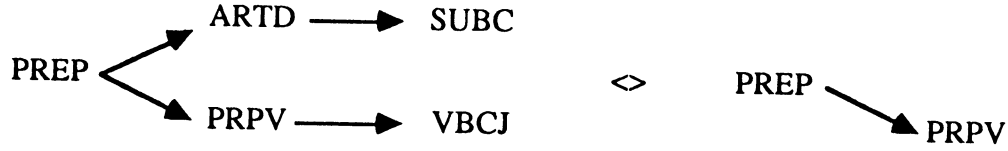
considérons la portion de phrase suivante : "... *les portions* ..." nous avons une ambiguïté classique :

ARTD —1→ SUBC

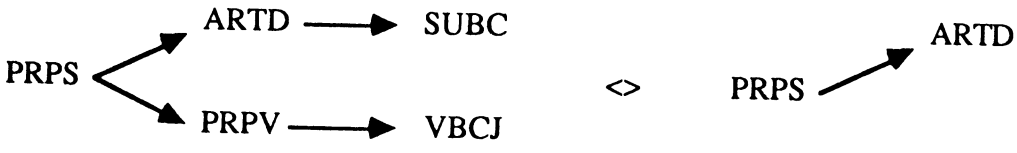
PRPV —2→ VBCJ

il n'existe pas une solution unique pour cette ambiguïté, cela va dépendre du contexte :

si la forme précédente est une préposition, on a toujours la solution n°1, ce qui peut se représenter par le schéma suivant :



si la forme précédente est un pronom personnel sujet, on a toujours la solution n°2, ce qui peut se représenter alors par le schéma suivant :



Toutes les ambiguïtés rencontrées ne permettent pas forcément la détermination d'un schéma, et c'est à l'utilisateur que revient la délicate responsabilité de cette détermination. Il doit pouvoir décider si le contexte rencontré caractérise une constante de résolution (ou de simplification). A cet effet, la simplicité du modèle linguistique utilisé et sa proximité avec les modèles classiques sont prépondérants.

Afin de laisser le plus de latitude possible à l'utilisateur (plus d'informations disponibles) pour la définition de la partie gauche d'un schéma, l'ensemble des "chemins pendants" résultants du processus de filtrage syntaxique est conservé jusqu'au processus d'application des schémas. Ceci permet de mieux caractériser les contextes des configurations d'ambiguïté, en disposant d'informations qui même si elles sont parasites, peuvent aider à discriminer ces configurations.

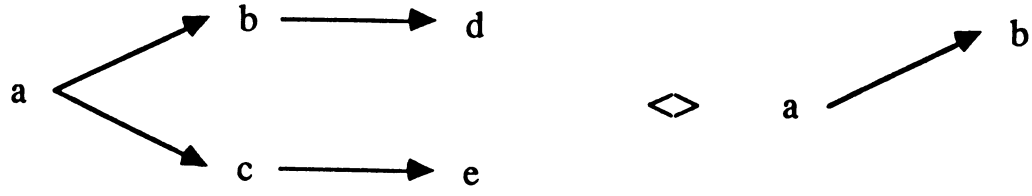
5.5.2. Détermination de la partie droite d'un schéma

La partie droite d'un schéma doit permettre la résolution ou la simplification de l'ambiguïté recensée en partie gauche. Afin de faciliter la mise en oeuvre, il est plus intéressant de donner en partie droite d'un schéma l'ensemble d'arcs de la partie gauche à retirer du réseau, plutôt que l'ensemble d'arcs à conserver; ces deux formulations sont équivalentes.

La partie gauche d'un schéma ne pouvant être qu'une partie de la portion de réseau ambiguë, il est important de ne retirer de cette portion que l'ensemble d'arcs nécessaire au solutionnement de l'ambiguïté reconnue, comme nous le montrons sur l'exemple ci-après.

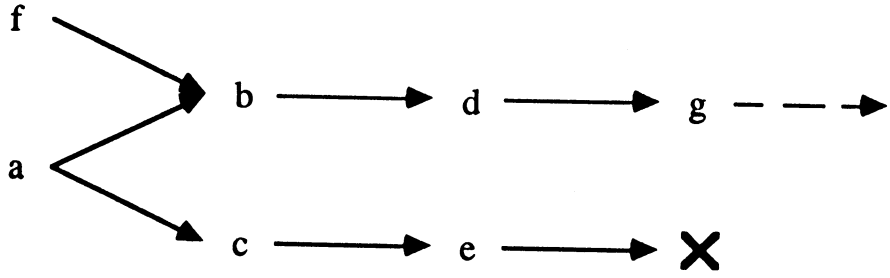
Exemple :

il suffit pour résoudre une ambiguïté dont la configuration graphique est du type décrite en partie gauche du schéma suivant, de réduire la première transition, en éliminant par exemple l'arc d'extrémité b :



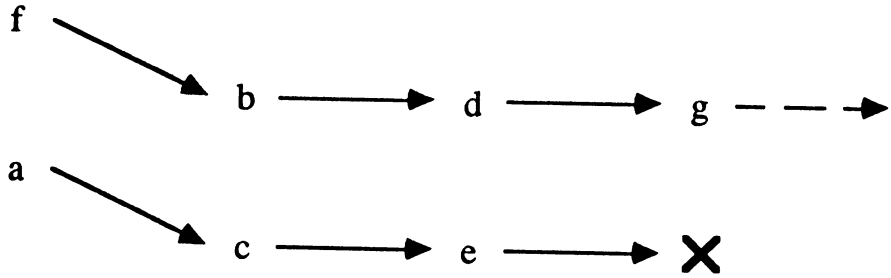
En effet si la transition correspondante du réseau ne comporte pas d'autre arc ayant comme extrémité b, ce sommet constituera l'origine d'un ou de plusieurs chemins pendants qui seront éliminés par le processus de nettoyage du réseau.

Cet aspect minimaliste dans la détermination des parties droites des schémas de résolution d'ambiguïté est essentiel pour garantir l'exactitude des résultats obtenus. Supposons que le schéma donné dans l'exemple précédent soit appliqué sur la portion de réseau suivante :



voie sans issue

Il est clair que la suppression de l'arc d'origine b, aurait conduit à l'élimination du chemin d'origine f et d'extrémité g, qui constitue pour cet exemple le seul chemin valide de cette portion de réseau, et par conséquence la seule interprétation valide de la portion de phrase correspondante. Le réseau résultat est alors :



voie sans issue

le chemin issu de "a" sera éliminé puisque "pendant", lors du nettoyage du réseau.

5.5.3. Cohérence du catalogue

Par cohérence du catalogue des schémas de résolution d'ambiguïté nous entendons :

- d'une part signaler les résolutions contradictoires apparentes, c'est-à-dire impliquant simultanément la suppression et la conservation d'un même ensemble d'arcs. C'est ce que nous appelons cohérence de la résolution, notion qui est développée en V.5.5.3.1;
- et d'autre part essayer de prévenir l'obtention de résultats incohérents tels que par exemple l'élimination de tous les arcs d'une transition du réseau; ce qui interromprait les chemins permettant l'interprétation continue de la portion de phrase correspondante, en créant un "trou". C'est ce que nous appelons cohérence du résultat qui est développée en V.5.5.3.2.

Avec la première stratégie mise en oeuvre (cf. V.5.4.1), le problème était simplement traité : lors de la définition des schémas par une interdiction de recouvrement suffixe des parties gauches, et par le choix de la stratégie d'application n'autorisant la validation que d'un seul schéma sur une portion de réseau donnée. On ne pouvait donc pas avoir de résolutions contradictoires, et par construction des schémas, les "trous" étaient évités. Ce n'est plus si simple, avec la possibilité d'application simultanée de plusieurs schémas.

Les schémas de résolution d'ambiguïté grammaticale sont de la forme :

$$S :: G <> D \quad \text{où :}$$

• G représente un ensemble d'arcs reconnus dans une portion du réseau analysé (séquence de transitions). On peut associer à G un prédicat g qui exprime les conditions d'existence des arcs a_i composants de G :

$$\text{soit } p_i = \exists\, a_i\,, \quad \text{on a alors} \quad g = p_1 \wedge \ldots\ldots \wedge p_n\,.$$

• D représente un sous-ensemble de G et définit les arcs à supprimer de G. On représente par le prédicat s_i le fait que l'on ait "$a_i \in G$ est à supprimer" et par $\overline{s_i}$ le fait que l'on ait "$a_i \in G$ est à conserver". On associe alors à D un prédicat $d = s_1 \wedge \ldots\ldots \wedge \overline{s_k} \wedge s_{k+1} \wedge \ldots\ldots \wedge \overline{s_n}$

A tout schéma S on peut associer l'implication logique :

$$g \longrightarrow d \equiv \overline{g} \vee d$$

en effet la table de vérité de l'implication nous donne :

g	d	g ⟶ d
V	V	V
V	F	F
F	V	V
F	F	V

son interprétation est la suivante :

-1) La 1ère ligne de la table correspond à la reconnaissance du schéma S et à son application.

-2) La 2ème ligne est une contradiction : la partie gauche est reconnue, mais la partie droite ne peut être appliquée, car elle correspond à une contradiction au sens de la logique. C'est le seul cas de contradiction dans la définition d'un schéma.

-3) La 3ème ligne correspond au cas où le prédicat d est vérifié sans que le schéma ait été reconnu; plusieurs schémas peuvent avoir concouru à la suppression des arcs de D. Ceci exprime que le prédicat g est une condition suffisante pour d.

-4) La 4ème ligne correspond au cas où le schéma n'est pas reconnu, et l'application d'autres schémas ne supprime pas tous les arcs de D.

La contradiction (cas 2) qui correspond à la vérification de g et à la non vérification de d, nous intéresse particulièrement pour la vérification de cohérence des schémas.

5.5.3.1. Cohérence de la résolution

Etant donné un ensemble S_1, ... ,S_n de schémas, vérifier leur cohérence consiste à :

a) vérifier qu'aucun d'eux ne comporte de contradiction; ce qui est vrai par définition et par construction des schémas (cf. V.5.1 et V.5.5.2). Cette vérification intrinsèque doit être effectuée lors de l'ajout d'un nouveau schéma.

b) vérifier qu'aucune conjonction des schémas ne conduit à une contradiction; l'ensemble S_1, ... ,S_n constitue un ensemble de propositions dont les parties gauches peuvent être simultanément vérifiée sur une portion de réseau, il faut alors que les propriétés impliquées soient cohérentes (c.a.d. n'engendrent pas de contradiction).

Ce niveau de cohérence s'analyse en considérant la conjonction de tous les schémas :

$$S_1 \wedge S_2 \wedge \ldots\ldots \wedge S_n.$$

Si aucun schéma ne correspond à une contradiction (cas a), il faut vérifier que leur conjonction n'en engendre pas non plus. Pour effectuer cette vérification, on peut se fonder sur les propriétés de la conjonction de deux implications logiques (opération associative), et réappliquer ensuite récursivement cette vérification sur le reste de la conjonction :

$$S_1 \wedge S_2 \wedge \ldots\ldots \wedge S_n = (\ldots(S_1 \wedge S_2) \wedge \ldots\ldots) \wedge S_n)$$

5.5.3.1.1. Propriétés du produit de deux implications

Soient $S_1 : g_1 \longrightarrow d_1$ et $S_2 : g_2 \longrightarrow d_2$
alors :

$$\begin{aligned} S_1 \wedge S_2 &\equiv (g_1 \longrightarrow d_1) \wedge (g_2 \longrightarrow d_2) \\ &\equiv (\overline{g_1} \vee d_1) \wedge (\overline{g_2} \vee d_2) \\ &\equiv \overline{g_1}\,\overline{g_2} \vee \overline{g_1}\,d_2 \vee \overline{g_2}\,d_1 \vee d_1 d_2 \\ &\equiv g_1 \vee g_2 \longrightarrow \overline{g_1}\,d_2 \vee \overline{g_2}\,d_1 \vee d_1 d_2 \quad (1) \end{aligned}$$

On peut en termes algorithmiques interpréter (1) par :

si $g_1 \vee g_2$ est vrai (g_1 ou g_2 ou les deux),
alors
- si g_1 est faux alors appliquer d_2 (car alors g_2 est vrai)
- si g_2 est faux alors appliquer d_1 (car alors g_1 est vrai)
- sinon (g_1 et g_2 sont vrais) appliquer d_1 et d_2

5.5.3.1.2. Combinaison de schémas

Dans le cas général, pour deux schémas de résolution d'ambiguïté on a :

$$g_1 \vee g_2 \longrightarrow \overline{g_1}\, d_2 \vee \overline{g_2}\, d_1 \vee d_1 d_2 \quad (1)$$

La contradiction apparait quand $g_1 \vee g_2$ est vrai (l'un ou l'autre ou les deux) et que la partie droite est fausse. Or si l'on suppose que par construction les deux schémas de départ ne comportent pas de contradiction, la contradiction engendrée par leur conjonction est décelable dans le cas particulier où g_1 et g_2 sont simultanément vrais. En effet :

- si g_1 est vrai et g_2 est faux alors (1) devient $g_1 \longrightarrow d_1$ qui ne comporte pas de contradiction;

- si g_2 est vrai et g_1 est faux alors (1) devient $g_2 \longrightarrow d_2$ qui ne comporte pas de contradiction;

- si g_1 et g_2 sont vrais, alors d_1 et d_2 sont également vrais, et (1) devenant $g_1 \vee g_2 \longrightarrow d_1 d_2$ ne peut comporter de contradiction que si $d_1 d_2$ est une contradiction.

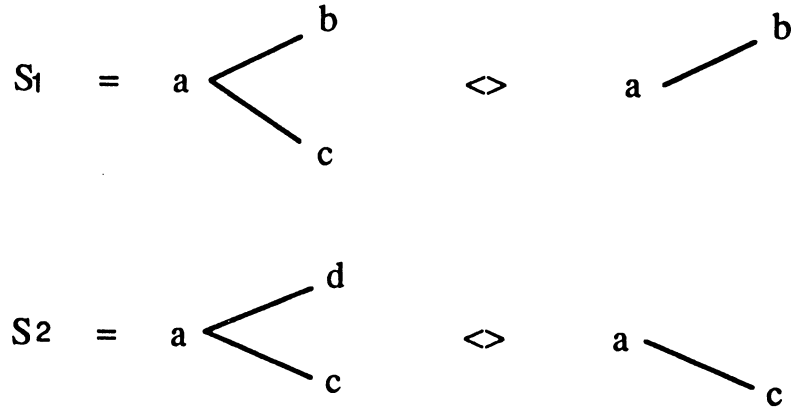
Donc, l'étude de la cohérence de la résolution de l'ensemble des schémas, repose sur l'étude de la cohérence de la résolution de chacun des schémas, et sur l'étude des contradictions pouvant résulter de leur combinaison.

Une contradiction logique $d_1 d_2$ résultant d'une combinaison, ne peut apparaître que lorsque simultanément on doit supprimer et conserver le même ensemble d'arcs. Or le caractère incrémentiel de la constitution du catalogue, qui permet entre autres de décrire une ambiguïté et son contexte de résolution à l'aide de plusieurs schémas, peut générer ce type d'incohérence logique qui ne correspond pas forcément à une incohérence de résolution.

Néanmoins, il est important que cette potentialité soit signalée lors de la création d'un schéma, de façon à pouvoir prendre conscience de la possibilité d'une résolution contradictoire, et donc de pouvoir éventuellement modifier les schémas concernés.

Exemple :

Soient les deux schémas suivants :



En appliquant la définition précédente, et en notant :

$p_1 : \exists$ ab , $p_2 : \exists$ ac et $p_3 : \exists$ ad

la définition des deux schémas se ramène à :

$$S_1 : \; p_1\, p_2 \longrightarrow s_1\, \overline{s_2} \equiv \overline{p_1} \vee \overline{p_2} \vee s_1\, \overline{s_2}$$

et

$$S_2 : \; p_2\, p_3 \longrightarrow s_2\, \overline{s_3} \equiv \overline{p_2} \vee \overline{p_3} \vee s_2\, \overline{s_3}$$

or si les deux schémas sont applicables en même temps, cela signifie d'après (1) que :

$$\begin{aligned} S_1 \wedge S_2 &\equiv p_1 \wedge p_2 \vee p_2 \wedge p_3 \longrightarrow s_1\, \overline{s_2}\; s_2\, \overline{s_3} \\ &\equiv p_1 \wedge p_2 \vee p_2 \wedge p_3 \longrightarrow \text{faux} \\ &\equiv \text{vrai} \longrightarrow \text{faux} \end{aligned}$$

on a donc une contradiction au sens de la logique, qui met en évidence une incohérence apparente des schémas S_1 et S_2.

5.5.3.2. Cohérence du résultat

Cette étude de cohérence du résultat a pour objectif de se prémunir contre des possibilités de création de "trou" dans le réseau, résultant d'une application simultanée de plusieurs schémas. Un "trou" étant l'élimination de tous les arcs

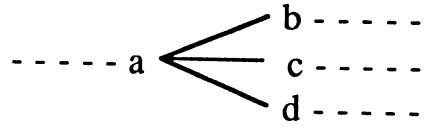
d'une transition du réseau analysé. En effet, si nous considérons deux schémas :

$$S_1 :: G_1 <> D_1 \quad \text{et} \quad S_2 :: G_2 <> D_2$$

leur application simultanée implique que les parties gauches G_1 et G_2 ont été reconnues sur la même portion de réseau. G_1 et G_2 comportent donc au moins une transition chacune, reconnue dans la même transition du réseau. Les parties droites D_1 et D_2 sont alors appliquées sur cette même portion de réseau. Or il se peut que l'ensemble résultant de l'union des arcs définis dans une des transitions correspondantes de G_1 et G_2 soit égal a l'ensemble résultant de l'union des arcs définis dans les transitions correspondantes des parties droites D_1 et D_2. Ce cas peut mener à la création d'un "trou", si la transition correspondante du réseau est composée de ce même ensemble d'arcs.

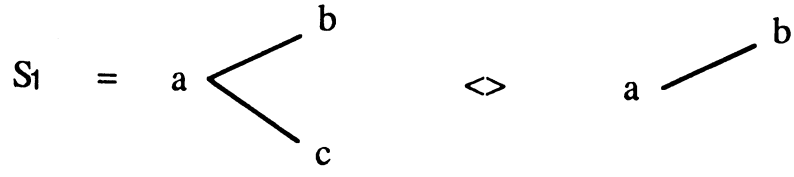
Exemple :

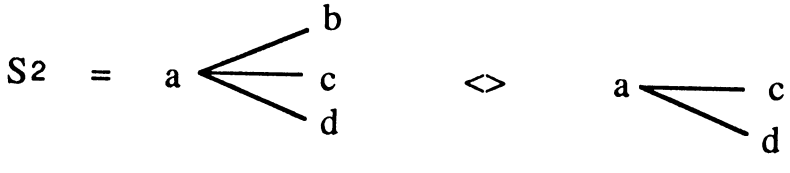
soit la portion de réseau constituée de la transition suivante :



T_i

et soient les deux schémas, sans contradiction, suivants :





leur application sur le réseau va supprimer tous les arcs de T_i et donc créer un "trou".

Ce type de problème n'est pas une incohérence intrinsèque : chacun des schémas concourant à l'élimination de tous les arcs de T_i, est correct. Seule la concomitance entre :

- d'une part l'égalité, entre l'ensemble d'arcs reconnus en partie gauche pour une transition donnée et l'ensemble correspondant des arcs à retirer (définis en partie droite des schémas),
- et d'autre part l'égalité de cet ensemble avec celui constituant la transition correspondante du réseau analysée,

provoque ce phénomène.

Si les deux schémas définis dans l'exemple précédent sont appliqués sur une transition T_j du réseau, comportant un arc de plus que T_i, on n'obtient pas de "trou". L'existence de "trou" dépend, en définitive, des propriétés du réseau et donc du langage analysé. C'est pour cette raison que l'on ne parlera que de possibilité d'incohérence du résultat.

En conséquence, nous avons étudié les possibilités d'applications simultanées de schémas qui peuvent engendrer une incohérence. Cette vérification est effectuée systématiquement lors de l'ajout d'un nouveau schéma dans le catalogue. L'utilisateur est donc averti lors de la création du schéma, il peut alors modifier ce nouveau schéma et/ou les schémas concernés, de manière à supprimer cette possibilité.

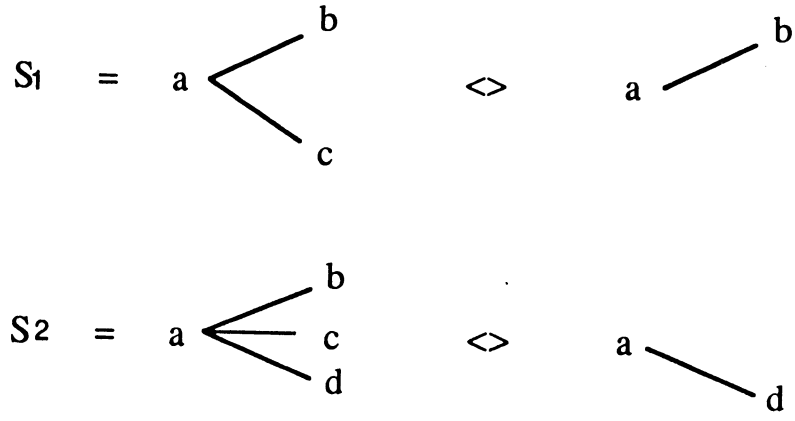
5.5.3.2.1. Saturation minimale d'une partie droite

La partie droite est définie par l'utilisateur lors de la création du schéma de résolution d'ambiguïté, en fonction de la configuration décrite en partie gauche. La saturation minimale de la partie droite d'un schéma consiste à compléter cette partie droite, avec les arcs provenant des parties droites des schémas inclus :

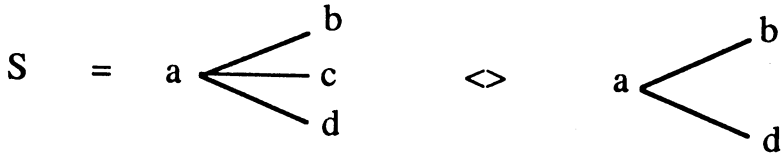
Un schéma S_1 est inclus dans un schéma S_2, si et seulement si la partie gauche G_1 de S_1 est incluse dans la partie G_2 gauche de S_2. L'inclusion de deux parties gauches est définie de manière analogue à l'inclusion d'un schéma de réseau dans un réseau (cf. IV.3.6.2.1, propriété 12)

Exemple :

soient les deux schémas suivants :



S_1 peut saturer S_2, car sa partie gauche est incluse dans celle de S_2, pour donner le schéma saturé minimal S :



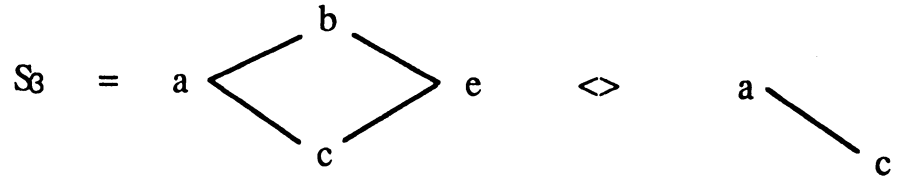
Un schéma saturé minimal, mémorise en partie droite l'ensemble d'arcs résultant de l'union des parties droites des schémas inclus, donc l'ensemble des arcs qui seront effectivement supprimés du réseau, lorsque la condition d'applicabilité du schéma aura été vérifiée, compte tenu des schémas inclus définis dans le catalogue.

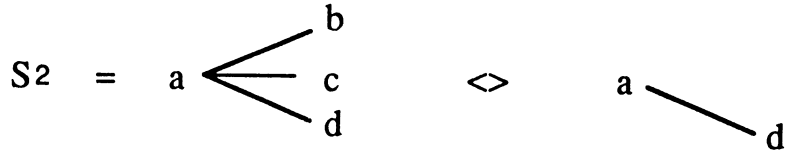
5.5.3.2.2. Saturation maximale d'une partie droite

On peut définir une saturation de la partie droite d'un schéma plus complète que la saturation minimale. Il suffit pour cela de comparer chaque transition de la partie gauche avec l'ensemble des transitions définies en partie gauche des autres schémas; et de compléter l'ensemble d'arcs à retirer de la partie droite avec les arcs retirés dans les autres schémas pour toutes les transitions incluses.

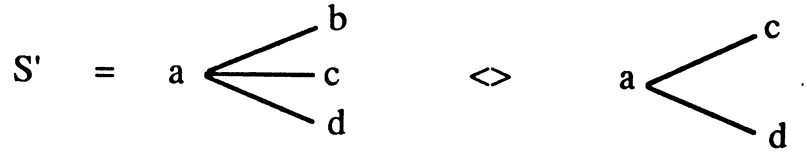
Exemple :

soient les deux schémas suivants :





S_3 peut saturer S_2 pour donner le schéma saturé maximal S' :



puisque la première transition de la partie gauche de S_3 est incluse dans la transition de la partie gauche de S_2.

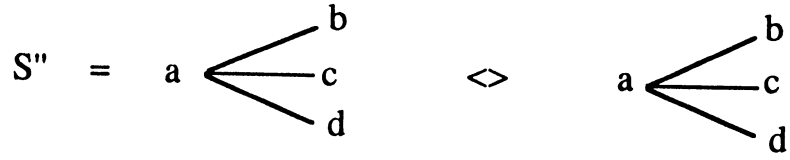
La partie droite d'un schéma saturé maximal contient tous les arcs présents en partie gauche, qui peuvent être supprimés lorsque la partie gauche du schéma est reconnue. La saturation maximale englobe la saturation minimale.

5.5.3.2.3. Etude des possibilités d'incohérence du résultat

Nous avons étudié les possibilités d'incohérence du résultat identifiables à partir des saturations minimales et maximales des parties droites des schémas de résolution d'ambiguïté. En effet lorsqu'une transition d'une partie droite contient le même ensemble d'arcs que la transition correspondante de la partie gauche, il y a risque de création d'un "trou", puisque tous les arcs reconnus sont supprimés. On dira alors que l'application du schéma concerné peut générer une incohérence de résultat.

Exemple :

si l'on considère les schémas S_1, S_2 et S_3 définis dans les deux exemples précédents, le schéma S_2 peut être saturé par les schémas S_1 et S_3, pour donner le schéma saturé maximal S" suivant :



Ce schéma comportant la même transition en partie gauche et en partie droite, l'application de S_2 pourra donc générer une incohérence de résultat.

Les conditions de réalisation d'un "trou" dont la possibilité est mise en évidence par la saturation d'une partie droite d'un schéma, outre l'égalité entre la transition éliminée de la partie gauche et la transition correspondante du réseau, sont :

- pour une saturation minimale : la reconnaissance de la partie gauche du schéma saturé;

- pour une saturation maximale : la reconnaissance de la partie gauche du schéma saturé, plus la reconnaissance des parties gauches des schémas participant à la saturation.

Les conditions de réalisation des "trous potentiels" détectés sont donc beaucoup plus exigeantes avec une saturation maximale qu'avec une saturation minimale. De plus la réalisation des saturations minimales des parties droites des schémas du catalogue est algorithmiquement simple (test d'inclusion); le maintien des schémas dans un état saturé l'est également :

Partant d'un catalogue initial vide, il suffit lors de chaque ajout, de saturer les schémas du catalogue avec le nouveau, et de saturer le nouveau avec ceux du catalogue.

5.5.3.3. Conclusion

Nous nous restreignons donc pour le maintien de la cohérence du catalogue des schémas de résolution d'ambiguïté à :

- vérifier que chaque schéma est construit conformément à la définition donnée en V.5.1;

- et à prévenir l'utilisateur des possibilités d'incohérence de résultat identifiables par une saturation minimale des parties droites des schémas.

Ce calcul s'effectuant lors de l'ajout des schémas, on pourrait envisager un traitement plus complet, les performances quantitatives de l'analyseur n'étant pas affectées.

5.6. Définition de classes de schémas de résolution d'ambiguïté

Afin de permettre une définition aisée de ces schémas, nous avons introduit des opérateurs qui permettent une définition concise de classes d'arcs, par des contraintes sur leurs sommets.

On représente par C l'ensemble des catégories grammaticales définies dans le modèle linguistique. Nous utilisons les quatre opérateurs de définition de classes d'arcs suivants :

-1) l'opérateur liste :

il permet de définir une classe de sommets par regroupement de plusieurs catégories grammaticales de C, il est matérialisé par les accolades { et }.

Exemple :

ici la liste représente un ensemble de sommets origine :

{ subc adjq vbpa } ---> vbcj

ce qui permet de définir l'ensemble des trois arcs suivants :

subc ---> vbcj
adjq ---> vbcj
vbpa ---> vbcj

-2) l'opérateur joker :

il correspond à une classe de sommets contenant toutes les catégories grammaticales de C. Il permet de définir une classe d'arcs dont seule l'origine ou l'extrémité est spécifiée. il est noté *.

Exemple :

ici l'utilisation de l'opérateur joker comme origine d'un arc,

* ---> vbcj

permet de représenter l'ensemble des arcs d'extrémité "vbcj".

-3) l'opérateur différent :

il définit une classe de sommets par exception. Il est noté ≠.

Exemple :

≠ subc ---> vbcj

représente l'ensemble des arcs d'extrémité "vbcj" et d'origine autre que "subc";

≠ { subc adjq } ---> vbcj

représente l'ensemble des arcs d'extrémité "vbcj" et d'origine autre que "subc" ou "adjq".

-4) l'opérateur non :

il définit une classe d'arcs par exception : il est noté ¬ .

Exemple :

¬ subc ---> vbcj

représente la classe des arcs définis sur C x C - (subc ---> vbcj);

¬ subc ---> *

représente la classe des arcs qui n'ont pas "subc" comme origine.

Les classes d'arcs ainsi définies permettent une factorisation de nombreux schémas et évitent une multiplication inutile qui alourdirait le processus d'application en multipliant les tests de reconnaissance dans le réseau.

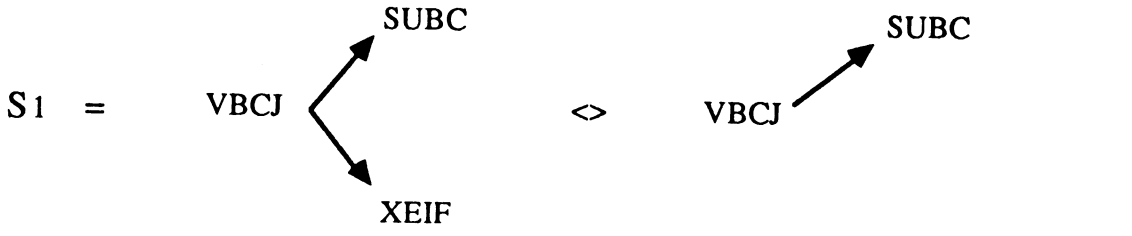
Leur interprétation algorithmique est immédiate :

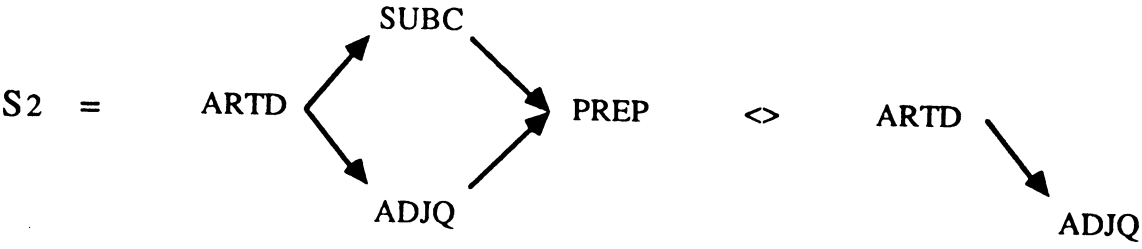
- en partie gauche, la reconnaissance d'un élément de la classe suffit;

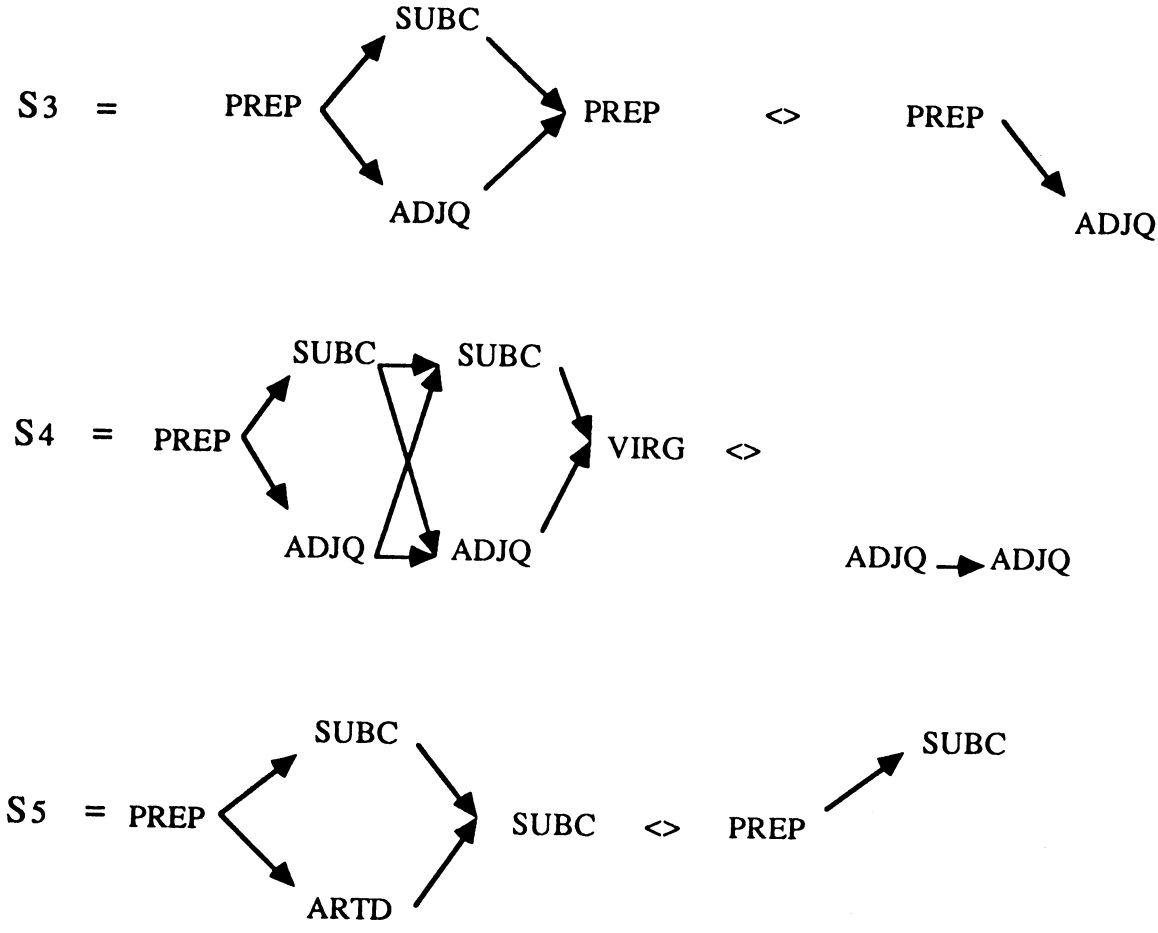
- en partie droite, tous les éléments de la classe présents dans la transition concernée du réseau sont éliminés par l'application du schéma.

5.7. Application des Schémas de Résolution d'Ambiguïté pour la phrase exemple

A titre d'exemple, nous donnons ci-dessous les schémas de résolution d'ambiguïté extraits du catalogue, qui sont utilisés sur le réseau obtenu à l'issue du filtrage syntaxique de la phrase exemple. Afin d'en simplifier la présentation, nous n'avons représenté dans la partie gauche de ces schémas, que les configurations ambiguës présentes dans l'exemple. Leur partie gauche peut être en réalité plus complexe, de manière à prendre en compte avec un même schéma plusieurs configurations d'ambiguïté de même nature (aboutissant au même schéma de résolution).







La figure ci-après représente l'exécution du processus d'application de ces schémas. Pour plus de clarté, cette figure ne comporte plus les chemins pendants issus du filtrage syntaxique.

Les arcs en pointillé correspondent aux arcs supprimés lors de l'application des schémas de résolution d'ambiguïté. Les étiquettes " **Si** " permettent de retrouver dans la liste donnée précédemment les schémas appliqués.

Comme pour la figure représentant l'exécution du processus de filtrage syntaxique, les chemins pendants issus de l'application des schémas sont étiquetés **CP** (flèches en gras), et les solutions isolées sont soulignées.

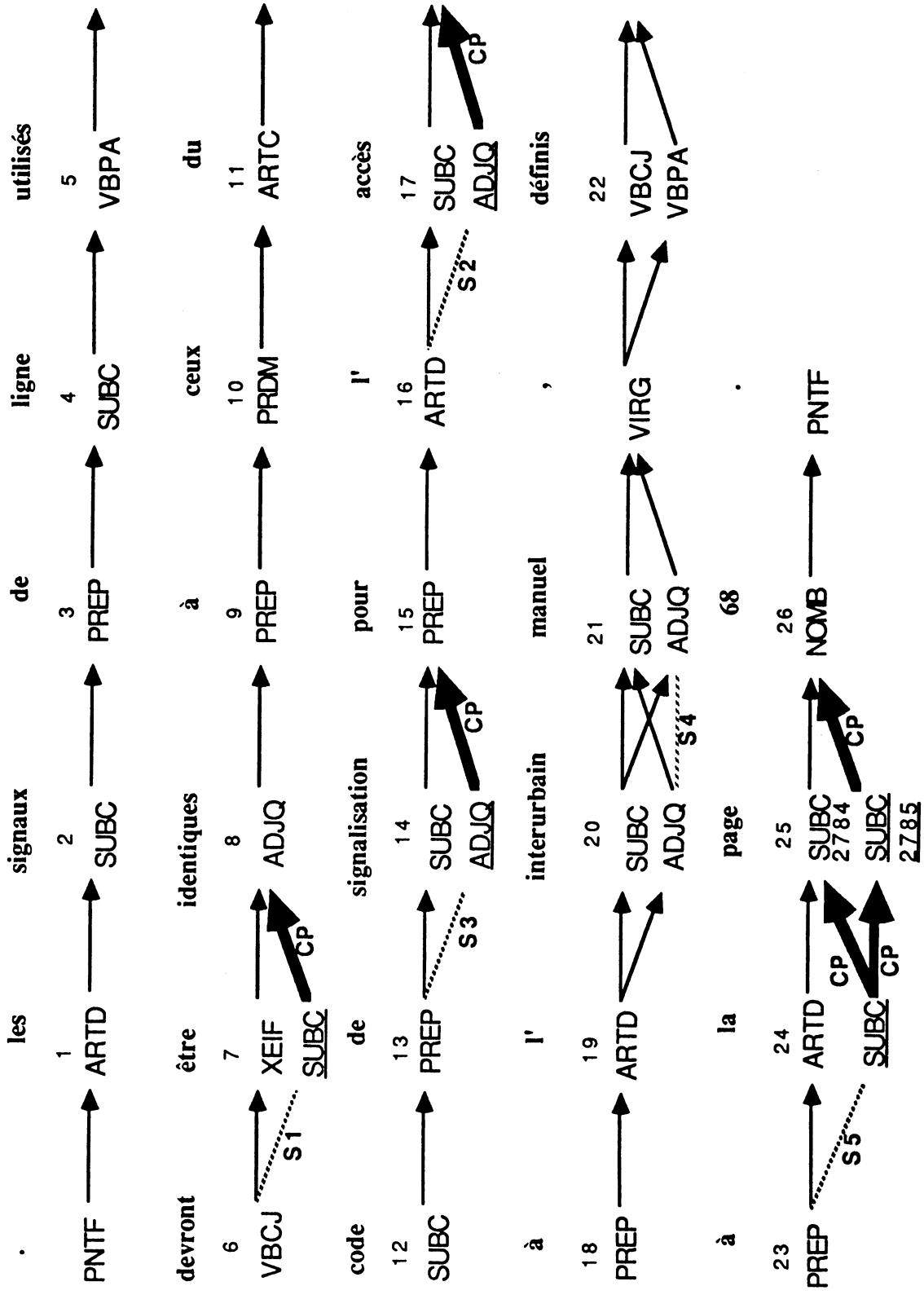


Figure 15 : *Application des schémas de résolution d'ambiguïté*

5.8. Nettoyage du réseau résultat

A l'issue du processus d'application des schémas de résolution d'ambiguïté, le réseau résultat comporte des chemins "pendants" et des solutions morphologiques isolées, résultants des simplifications effectuées sur le réseau initial.

Les chemins "pendants" correspondent à des impasses pour les tentatives d'interprétation du texte analysé (cf. IV.3.5), et doivent donc être éliminés avant l'exécution du processus d'extraction des Groupes Conceptuels. De même que les solutions morphologiques isolées, puisqu'elles ne font partie d'aucun chemin et qu'elles ne peuvent donc pas participer à l'interprétation d'une portion de texte.

A l'issue de la résolution des ambiguïtés grammaticales répertoriées à l'aide des schémas, et de l'élimination des chemins pendants, le nombre de chemins correspondant à autant d'interprétations potentielles de la phrase exemple n'est plus que de 6. Il reste en effet trois formes ambiguës admettant chacune deux solutions morphologiques : ce sont les formes n° 20, 21 et 22.

Le réseau résultat obtenu est un réseau simplifié. Les portions de réseau linéaires obtenues sont des réseaux linéaires maximum. Nous pouvons considérer ce réseau résultat $R_{1,26}$, comme la concaténation de deux réseaux linéaires maximums et d'un réseau ambigu, ce qui peut s'écrire (cf. IV.3.5.1.2 remarque 10) :

$$R_{1,26} = RM_{1,19} + RA_{19,23} + RM_{23,26}$$

Ce sont les portions de réseaux linéaires maximums de ce type, qui constituent ensuite l'entrée du processus d'extraction des Groupes Conceptuels.

La figure ci-après permet de visualiser ce réseau résultat simplifié :

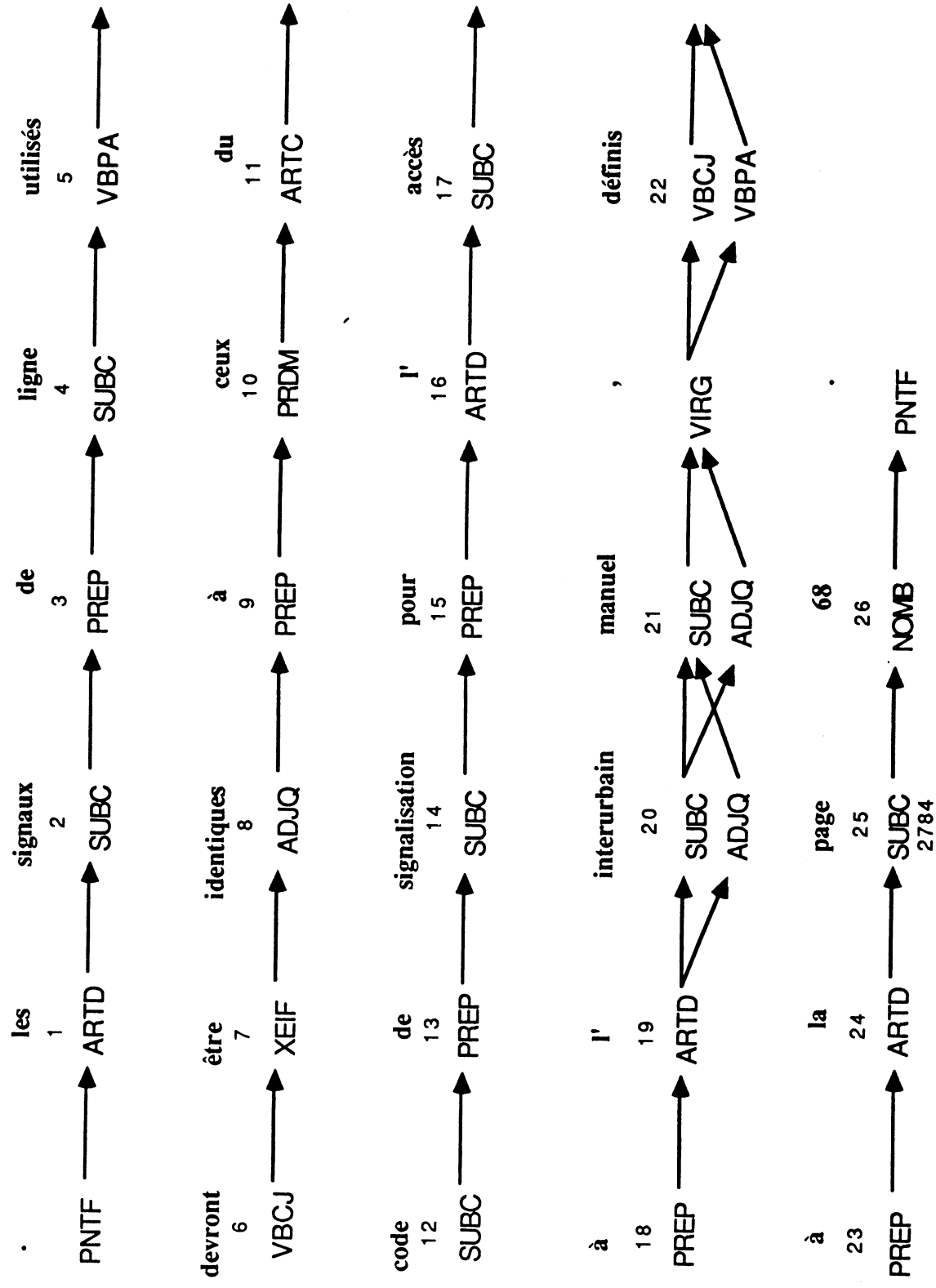


Figure 16 : *Réseau simplifié*

5.9. Conclusion sur la résolution des ambiguïtés grammaticales

Ce processus de résolution d'ambiguïté grammaticale, conformément aux objectifs dévolus à l'analyse développée, permet de résoudre par une utilisation de mécanismes dont la mise en oeuvre est simple et efficace, un certain nombre des ambiguïtés persistantes après le filtrage positionnel et grammatical.

Il repose sur une représentation des connaissances constituées par les constantes de résolution des ambiguïtés, dans un formalisme identique à celui dans lequel sont manipulées les données (notion de réseau).

Il présente une grande souplesse de définition qui permet en particulier un ciblage des ambiguïtés à résoudre. Le caractère incrémentiel de la constitution du catalogue des schémas permet de procéder par étapes et ne nécessite pas la définition d'un système complet pour être utilisé. Nous nous sommes, par exemple, particulièrement intéressés pour notre application, aux résolutions des ambiguïtés utiles pour la reconnaissance des syntagmes nominaux.

6. Le traitement des formes incohérentes, catégorisation automatique

La catégorisation automatique intervient dans le processus d'analyse lors de la rencontre d'une forme incohérente. Nous rappelons qu'une forme incohérente est une forme dont la transition associée est incohérente (cf. IV.3.3.2 -b). Une forme incohérente constitue donc une discontinuité dans l'interprétation d'une portion de texte.

Deux cas peuvent se produire :

- la forme analysée est inconnue (son ensemble solution est vide),
- la forme analysée est incohérente gauche ou droite (son ensemble solution n'est pas vide, mais est peut-être incomplet).

L'objectif de la catégorisation automatique est de déterminer pour une forme incohérente un ensemble solution potentiel (cas de la forme inconnue), ou de compléter l'ensemble solution fourni par l'analyse morphologique (cas d'une forme incohérente gauche ou droite incomplètement interprétée). On rappelle qu'un ensemble solution est constitué de triplets comprenant : un représentant d'unité lexicale, une catégorie grammaticale et un ensemble de valeurs grammaticales.

Le traitement des formes incohérentes, que nous appelons aussi catégorisation automatique, nécessite donc la détermination :

-1) des catégories grammaticales potentielles,

-2) des variables grammaticales associées à ces catégories,

-3) d'un représentant d'unité lexicale pour chaque couple :
catégorie - ensemble de valeurs grammaticales .

6.1. Détermination du contenu initial du dictionnaire

En conséquence, nous avons répertorié tout d'abord, l'ensemble des classes fermées du français, avec lesquelles nous avons procédé à l'initialisation du dictionnaire (cf. IV.2.1). Puis nous y avons ajouté l'ensemble des irrégularités que nous avons pu recenser, principalement à partir de deux ouvrages " *Le nouveau Bescherelle : 1.l'art de conjuguer* " [BESC 80], et "*Le bon usage* " [GREV 80].

Toutes les formes constituant ces deux ensembles, ont été décomposées manuellement (et ont servi à cette occasion à la détermination des tables de désinences et à la composition des modèles morphologiques). Les radicaux résultant de cette décomposition, et les renseignements linguistiques associés, ont été consignés dans le dictionnaire d'analyse.

Nous avons constaté alors, que la plupart des mots les plus fréquents de la langue se retrouvent parmi les éléments des classes fermées et de l'ensemble des irrégularités. Ce qui signifie qu'à partir de cette initialisation, on peut reconnaître déjà, la grande majorité des mots utilisés dans la rédaction de textes écrits en français. Afin de renforcer ce phénomène, et de limiter l'apparition de mots inconnus du système, quelques éléments des classes ouvertes, relativement fréquents, ont également été chargés lors de la phase d'initialisation. Nous avons utilisé pour cela, outre les ouvrages précités, "*Le dictionnaire des fréquences* " constitué sur un très gros corpus de textes représentatifs de la littérature française [TLF 71].

Cet ensemble de données initiales du dictionnaire permet d'effectuer deux inférences importantes pour le traitement des mots inconnus :

- tout mot inconnu fait partie au moins d'une classe ouverte (possibilité d'homographie),
- tout mot inconnu a un comportement grammatical régulier.

C'est à partir de ces deux constatations que nous avons pu établir un processus de catégorisation automatique des formes incohérentes, qui est à la base du processus d'enrichissement automatique du vocabulaire (cf. V.7).

Néanmoins, cette initialisation du dictionnaire d'analyse n'est pas suffisante pour obtenir un fonctionnement optimum de l'analyseur morphologique, même si celui-ci est couplé avec un processus d'enrichissement automatique

6.2. Détermination des catégories grammaticales potentielles

Cette catégorisation repose sur la distinction que nous avons établie entre les classes fermées (dont les listes exhaustives d'éléments sont connues et répertoriées), et les classes ouvertes. Nous avons vu précédemment que cette distinction nous permet de déduire pour une forme inconnue ou pour une forme incohérente gauche ou droite, que l'ensemble solution des catégories grammaticales hors contexte qui peut lui être attribué, est un sous-ensemble de l'ensemble des catégories grammaticales des classes ouvertes qui sont au nombre de neuf dans le modèle linguistique défini cf. IV.2.2) et que nous rappelons ici :

SUBC substantifs communs,
SUBP substantifs propres,
ADJQ adjectifs qualificatifs,
ADVB adverbes,
VBIF verbes à l'infinitif,
VBPA verbes au participe passé,
VBPR verbes au participe présent,
VBCJ verbes conjugués,
ABRV abréviations et sigles.

Nous pouvons préciser un peu plus ce sous-ensemble, en constatant que pour certaines de ces classes, l'ensemble des terminaisons possibles pour leurs éléments, est très restreint :

- pour la catégorie ADVB (adverbes), seule la terminaison " *ment* " est possible (les adverbes dont la terminaison est différente ont été recensés et consignés lors de l'initialisation du dictionnaire d'analyse),

- pour la catégorie VBIF (verbes à l'infinitif), seules les terminaisons " *er* " et " *ir* " sont possibles (les verbes du troisième groupe de conjugaison permettant d'autres terminaisons font partie des irrégularités, et ont donc également été consignés lors de l'initialisation du dictionnaire d'analyse),

- pour la catégorie VBPR (verbes au participe présent), seule la terminaison " *ant* " est possible,

- pour la catégorie VBPA (verbes au participe passé), seules les terminaisons " *é* ", " *ée* ", " *ées* ", " *és* ", " *i* ", " *ie* ", " *ies* " et " *is* " sont possibles pour les premier et deuxième groupes de conjugaison,

- il en est de même pour la catégorie VBCJ (verbes conjugués), où la liste est un peu plus longue (ensemble des désinences des premier et deuxième groupes de conjugaison) : " *e* ", " *es* ", " *ez* ", etc...

On peut remarquer que pour les quatre catégories ADVB, VBIF, VBPR et VBPA, les ensembles de terminaisons possibles sont disjoints, ce qui signifie que deux de ces catégories ne pourront jamais être déterminées en même temps.

Néanmoins, le cas le plus fréquent reste celui des terminaisons communes à plusieurs catégories.

Exemple :

la chaîne "*ment* " peut être la terminaison d'un adverbe ADVB, d'un verbe conjugué VBCJ, d'un substantif commun SUBC ou d'un adjectif qualificatif.

Les catégories SUBP (nom propre) et ABRV (abréviation et sigle) dont le comportement syntaxique est similaire à celui des substantifs communs (cf.IV.2.2), seront déterminées en fonction de particularités de typologie autres que les terminaisons, et qui sont :

- première lettre en majuscule pour les noms propres,
- et par exemple toutes les lettres majuscules pour les abréviations et sigles.

La reconnaissance des formes admettant ces catégorisations est effectuée par des procédures spécifiques du processus de lecture des formes en entrée comme nous l'avons vu précédemment (cf. V.3.2.3).

Nous n'avons pas établi de restrictions particulières pour les terminaisons des substantifs communs et des adjectifs qualificatifs (SUBC et ADJQ), considérant a priori que toute forme inconnue peut appartenir à ces catégories. L'établissement de listes de terminaisons pertinentes nécessiterait un travail important pour pouvoir être efficace, car outre la difficulté liée à la nécessaire exhaustivité de ces listes, ce travail requiert une très bonne connaissance de la langue. De plus l'acquisition permanente de nouveaux mots d'origine étrangère commanderait des mises à jour fréquentes de ces listes.

Nous avons donc établi une partition de l'ensemble des catégories grammaticales des classes ouvertes (hormis les catégories SUBP et ABRV) basée sur la compatibilité de certaines terminaisons avec plusieurs catégories. L'ensemble de ces partitions ayant comme intersection commune les deux catégories grammaticales SUBC et ADJQ, qui sont systématiquement attribuées.

En conséquence, lors de la rencontre d'une forme inconnue ou d'une forme incomplète, la terminaison de cette forme déterminera la partition qui pourra lui être attribuée hors contexte.

Exemple :

Supposons que la forme " *établissement* " soit une forme inconnue, l'ensemble des catégories potentielles qui pourra lui être attribué automatiquement sera :

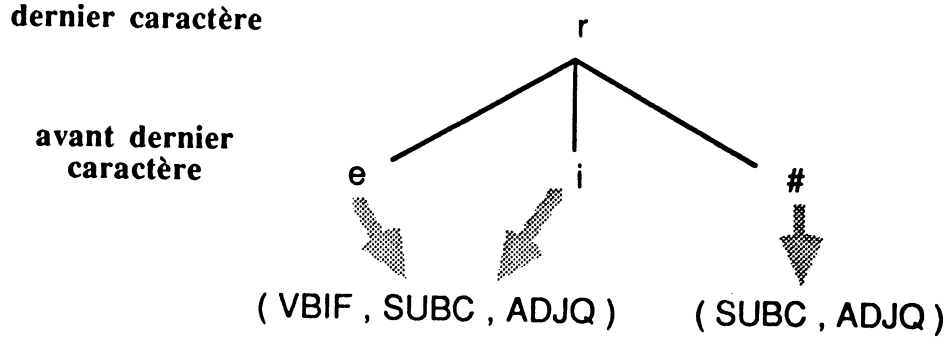
ADVB (terminaison en " *ment* ")
VBCJ (terminaison en " *ent* ")
SUBC
ADJQ

Cet ensemble comprend bien la solution réelle qui est SUBC.

On peut représenter les déterminations des ensembles de catégories grammaticales potentielles par des arbres de décision constitués à partir des terminaisons possibles (il existe un arbre de décision pour chaque dernier caractère possible). Les feuilles de ces arbres correspondent aux partitions de l'ensemble des catégories grammaticales des classes ouvertes déterminées.

Exemple :

l'arbre de décision permettant de reconnaître les verbes à l'infinitif (ainsi que les substantifs et les adjectifs) est :



(le caractère '#', signifiant tout autre caractère)

Une fois les catégories grammaticales potentielles déterminées, on connait les variables grammaticales à instancier.

6.3. Détermination des valeurs des variables grammaticales

Le problème de cette détermination ne se pose que pour les catégories SUBC, ADJQ, VBCJ et VBPA, auxquelles sont associées des variables grammaticales. Nous pouvons discerner pour ces catégories, le cas des verbes pour lesquels les ensembles de terminaisons possibles sont complets, du cas des substantifs et des adjectifs.

6.3.1. Cas des verbes

Pour les deux catégories grammaticales VBCJ et VBPA, le problème est résolu par la donnée des terminaisons possibles. Ces terminaisons correspondant aux désinences verbales qui, comme on l'a vu précédemment (cf. la présentation des modèles morphologiques), ont leur ensemble de valeurs grammaticales propres associé.

Exemple :

la terminaison " *is* " commune aux deux catégories donnera :

------> ({ 1ère, 2ème } , singulier) pour les VBCJ,

------> (masculin , pluriel) pour les VBPA.

L'attribution des valeurs grammaticales pour la catégorie VBCJ sera complétée pour les variables MODE et TEMPS, en fonction des diverses désinences des premier et deuxième groupes de conjugaison admettant comme terminaison " *is* " : (" *ais* ", " *erais* ", " *issais* ", " *irais* ", " *is* "). Ces valeurs étant factorisées au niveau des tables contenant ces désinences (cf. V.3.3.1).

Exemple :

avec la désinence " *issais* " on obtient les quatre possibilités suivantes :

(indicatif, imparfait) pour *"issais"*

(indicatif, passé simple) pour *"ais"*

(impératif, présent) et (indicatif, présent) pour *"is"*

Ce sont donc, les valeurs grammaticales associées aux désinences lors de leur définition et de leur organisation sous forme de tables, qui affectent directement les variables grammaticales lors de la détermination d'une catégorie VBCJ, ou VBPA.

6.3.2. Cas des substantifs et des adjectifs

Pour les deux catégories grammaticales SUBC et ADJQ, il en va autrement, d'autant que, comme nous venons de le voir, ces deux catégories sont systématiquement possibles. Les variables grammaticales associées à ces deux catégories sont GNR et NBR (genre et nombre).

La détermination de la valeur exacte de la variable NBR est un peu plus facile que pour la variable GNR, du fait qu'en français il n'existe que peu de façons de marquer le pluriel d'un nom ou d'un adjectif qualificatif. Par contre si certaines règles permettant de reconnaître le genre des noms et des adjectifs qualificatifs existent, elles comportent souvent des exceptions difficilement énumérables (sans parler des différences pouvant exister entre les différentes institutions de référence : Académie française, Littré, etc...).

En conséquence, nous avons effectué un certain nombre de choix qui, s'ils sont contestables, ont le mérite de nous permettre quelques simplifications (cf. annexe 5). Une étude plus systématique nécessiterait le concours de linguistes, et ne constitue pas l'objet de notre travail.

Nous avons procédé de la manière suivante : nous avons considéré manuellement, à partir des arbres de décision qui permettent de définir les ensembles de catégories grammaticales potentielles, tous les chemins aboutissant à des ensembles contenant les catégories SUBC et ADJQ. A partir de ces chemins correspondant aux terminaisons, on a essayé de déterminer les valeurs des variables GNR et NBR.

Quand il ne nous "paraissait" pas possible de déterminer une valeur précise, nous avons attribué les valeurs "doubles" :

{ masculin, féminin } pour la variable GNR,
{ singulier, pluriel } pour la variable NBR.

Si les affectations de valeurs grammaticales ne sont pas toujours précises (utilisation possible de valeurs doubles), elles constituent au pire un sur-ensemble de l'ensemble des valeurs "exactes" attribuables hors contexte. Ainsi

nous respectons le principe énoncé précédemment (cf. IV.3.2.2), qui prévoit que l'ensemble solution S, attribué à une forme inconnue par le processus de catégorisation automatique soit un sur-ensemble (autant que faire se peut, le plus petit sur-ensemble possible) de l'ensemble solution hors contexte, SM, déterminable dans le modèle linguistique.

6.4. Détermination du représentant d'unité lexicale

La détermination du représentant d'unité lexicale va s'effectuer à partir de la morphologie de la forme concernée, et en fonction du résultat des déterminations de la catégorie grammaticale et de l'ensemble de valeurs grammaticales associé.

Il s'agit en fait d'une normalisation de la forme considérée (cf. IV.3.1.-i) en choisissant lorsque c'est possible :

- la forme au masculin-singulier pour les adjectifs qualificatifs ADJQ et les participes passés VBPA,

- la forme au singulier pour les substantifs communs SUBC,

- la forme à l'infinitif pour les verbes VBCJ et VBIF,

- la forme telle qu'elle est rencontrée pour les autres catégories SUBP, ADVB, VBPR, ABRV.

C'est cette détermination du représentant d'unité lexicale qui va conditionner la possibilité d'acquisition automatique de l'unité lexicale. Si l'on peut déduire sans ambiguïté ce représentant et que de plus la solution morphologique est validée par l'analyse (cf. V.7.2), alors l'unité lexicale sera acquise automatiquement dans son intégralité : toutes les formes possibles seront prises en compte.

Le problème de la détermination de ce représentant n'est pas simple : il s'agit de décomposer la forme concernée en un couple racine + désinence afin de pouvoir recomposer le représentant en utilisant les critères que l'on vient d'énoncer. Les choix effectués lors de l'initialisation du dictionnaire pour la consignation des irrégularités grammaticales (cf. V.6.1), et lors de la morphologie pour le non traitement des préfixes (cf. V.3.5) nous permettent d'en limiter la complexité.

Nous allons expliciter notre démarche en étudiant le cas des substantifs communs.

Tout d'abord, il faut partir du principe que les formes au masculin et au féminin d'un substantif constituent deux unités lexicales différentes. Tout substantif n'admet pas forcément les deux genres, et dans certains cas les racines peuvent être différentes. Il est donc impossible sur des critères uniquement morphologiques, sans disposer d'informations complémentaires de nature sémantique ou même pragmatique, de pouvoir effectuer un regroupement en une seule unité lexicale sans tomber dans les travers consistant par exemple à considérer que le substantif "*porte* " est la forme au féminin du substantif "*port* ".

Cela étant précisé, et sans nous préoccuper du genre, la forme normalisée des substantifs communs doit être la forme au singulier lorsque c'est possible. Pour déterminer ce singulier et le pluriel dérivé, on va considérer que la forme concernée a un comportement grammatical régulier puisque toutes les irrégularités sont consignées.

De fait, pour déterminer le pluriel à partir de la forme au singulier, il suffit d'appliquer les règles suivantes :

-1) toute forme se terminant par "*s* ", "*x* " ou "*z* " au singulier ne change pas au pluriel,

-2) toute forme se terminant par "*al* " au singulier fait son pluriel en "*aux* ",

-3) toute forme se terminant par "*au* ", "*eau* " ou "*eu* " au singulier prend un "*x* " au pluriel,

-4) toute autre forme au singulier prend un "*s* " au pluriel.

Par contre la détermination du singulier à partir du pluriel n'est pas toujours possible, seuls les cas suivants qui sont loin d'être les plus nombreux, peuvent être traités de manière univoque :

-1) toute forme se terminant par "*z* " ne change pas au singulier,

-2) toute forme se terminant par "*eaux* " ou "*eux* " perd le "*x* " au singulier,

-3) toute forme se terminant par "*x* ", exceptées celles se terminant par *"aux* " ou *"eux* ", ne changent pas au singulier.

Ces considérations induisent donc pour la détermination du représentant d'unité lexicale d'un substantif commun la démarche suivante :

-1) si la forme rencontrée est au pluriel et ne rentre pas dans les trois cas que nous venons d'exposer, on ne peut pas déterminer la forme au singulier. On conserve donc la forme pluriel qui en cas d'acquisition automatique constituera le représentant jusqu'à la rencontre du singulier.

-2) si la forme rencontrée est au singulier, on n'a pas de difficulté pour déterminer le pluriel. Si ce pluriel existe déjà tout seul (cas précédent), on doit lui retirer son rôle de représentant.

-3) si le nombre de la forme rencontrée est indéterminé, on conserve cette ambiguïté jusqu'à la rencontre d'un des deux cas précédents.

Ce fonctionnement pourrait être considérablement amélioré par la connaissance de listes exhaustives telles que par exemple celle des substantifs communs se terminant par un "*s* " au singulier et au pluriel. Là encore l'établissement de telles listes incombe à la communauté des linguistes.

6.5. Catégorisation des formes inconnues

La rencontre d'une forme inconnue dans un texte, lors de l'analyse morphologique, active systématiquement le processus de catégorisation automatique que nous venons de décrire afin de déterminer pour cette forme un ensemble solution hors contexte S.

Une fois cet ensemble déterminé, la forme est numérotée et stockée temporairement (pendant l'analyse de la portion de texte en cours) dans une table des formes inconnues avec son ensemble solution hors contexte S, de manière à pouvoir l'identifier lors de la rencontre d'une nouvelle occurrence dans la même portion de texte (le processus de catégorisation automatique n'étant pas réactivé dans ce cas là).

Le filtrage positionnel et grammatical (cf. V.4) permet ensuite de vérifier la cohérence des interprétations de cet ensemble solution avec le contexte de la forme traitée (ce contexte est constitué des interprétations de la forme

précédente et de la forme suivante). Il en résulte un ensemble solution éventuellement réduit par rapport à l'ensemble solution hors contexte déterminé. Ce sont les interprétations de cet ensemble solution réduit qui sont validées dans la table des formes inconnues, de manière à en permettre ensuite la consignation dans le dictionnaire. Chaque interprétation ainsi validée reçoit un numéro d'unité lexicale, qui permettra de l'identifier par la suite.

6.6. Catégorisation des formes "incomplètes"

Nous avons vu au chapitre précédent (cf. IV.3.4.1 Remarque 8) que l'impossibilité d'obtenir une interprétation cohérente d'une portion de texte, outre la rencontre d'une forme inconnue, pouvait avoir deux origines :

- soit le texte d'entrée est incorrect, ce que nous excluons a priori,
- soit il existe au moins une forme pour laquelle l' analyse morphologique est incomplète. C'est ce cas que nous considérons en premier lieu.

Si à l'issue du traitement d'une forme potentiellement incomplète, la transition associée demeure toujours incohérente, nous en déduirons une contradiction entre le texte analysé et notre modèle d'analyse, et nous pourrons conclure soit à une incorrection du texte en entrée, soit à un défaut de notre démarche.

Nous allons définir ce que nous appelons forme potentiellement incomplète, car toute forme incohérente n'est pas forcément une forme incomplète.

Forme potentiellement incomplète

> Une forme potentiellement incomplète (par simplification on dira forme incomplète) est une forme incohérente gauche ou droite, ayant été répertoriée à partir du processus d'enrichissement automatique du vocabulaire.

Cela signifie que, lors de la rencontre de la première occurrence d'une forme (ayant permis son acquisition automatique), seule une partie des interprétations potentielles hors contexte est consignée. L'attribution automatique des solutions morphologiques pour une forme inconnue précède en effet le filtrage syntaxique, ce qui peut produire l'élimination de certaines

de ces solutions. Or lors de l'acquisition proprement dite de cette nouvelle forme, seules les solutions validées (et donc présentes) sont enregistrées.

Cela ne peut être le cas des formes provenant de l'ensemble des données initiales du dictionnaire (classes fermées, irrégularités, éléments les plus fréquents des classes ouvertes), puisque nous avons fait l'hypothèse que ces formes sont consignées avec un ensemble solution complet, construit manuellement. En effet, cette étude manuelle est rentable, puisque cet ensemble de données initiales correspond à un sous-ensemble relativement restreint du vocabulaire français, qui couvre à lui seul une part importante des formes d'un texte.

On peut distinguer à l'aide des informations contenues dans le dictionnaire d'analyse les formes acquises par le processus d'enrichissement automatique, des formes issues de l'ensemble des données initiales. Ainsi nous ne sommes pas conduit à rechercher pour une forme complète d'hypothétiques solutions supplémentaires.

Le traitement d'une forme incomplète consiste donc à retrouver parmi ces solutions morphologiques disparues (qui n'ont pas été validées pour l'enrichissement automatique), celle ou celles permettant de rendre cohérente la transition associée à cette forme. Il va de soi que les nouvelles solutions, une fois validées, sont consignées d'abord dans la table des formes inconnues, puis dans le dictionnaire conformément au déroulement du processus d'enrichissement que nous présentons plus avant (cf. V.7).

6.7. Conclusion

Au cours de l'analyse, l'exécution de ce processus de catégorisation automatique peut intervenir a deux moments différents :

-1) soit pendant le processus d'analyse morphologique, lors de la rencontre d'une forme inconnue,

-2) soit au cours de l'exécution du processus de filtrage syntaxique, lors de la rencontre d'une forme incohérente incomplète.

Dans ce dernier cas, le filtrage syntaxique doit être réappliqué pour cette forme, après avoir complété son ensemble solution.

Ce processus de catégorisation automatique peut interférer, en présence d'une forme incomplète, avec les processus d'analyse morphologique et de filtrage syntaxique. Nous nous sommes par contre interdit de le faire intervenir pendant ou à l'issue de l'exécution du processus de résolution des ambiguïtés grammaticales.

6.8. Catégorisation automatique pour la phrase exemple

Sur notre exemple, les trois formes inconnues du système vont être interprétées de la façon suivante :

signalisation	---> (	SUBC	MAS,FEM SIN)
	---> (	ADJQ	MAS,FEM SIN)
accès	---> (	SUBC	MAS,FEM SIN,PLU)
	---> (	ADJQ	MAS,FEM SIN,PLU)
interurbain	---> (	SUBC	MAS,FEM SIN)
	---> (	ADJQ	MAS,FEM SIN)

Ce résultat, qui peut être considéré comme une approximation du résultat souhaité, permet une poursuite de l'analyse sans intervention extérieure (de l'utilisateur par exemple). Pour que cette poursuite ait lieu dans des conditions satisfaisantes, il est important que chaque ensemble solution constituant ce résultat "approximatif", contienne la ou les solutions morphologiques exactes, ou du moins une ou des solutions morphologiques englobantes (cf. IV.3.2.1). Ce qui est le cas pour les ensembles solutions associés aux quatre formes inconnues rencontrées dans la phrase analysée.

Exemple :

La solution morphologique exacte pour " *signalisation* " est :

(SUBC FEM SIN)

cette solution se retrouve dans la première solution morphologique potentielle proposée :

(SUBC MAS,FEM SIN)

En effet, la valeur double de la variable "genre", {MAS,FEM}, contient la valeur de cette variable dans la solution morphologique exacte { FEM }.

7. Enrichissement automatique du vocabulaire

Le processus d'enrichissement automatique du vocabulaire constitue en partie une solution aux problèmes posés par la rencontre en cours d'analyse de mots inconnus, c'est-à-dire de mots nouveaux pour le système.

Ce mécanisme permet dans un certain nombre de cas " d'apprendre " un mot nouveau, et de déduire de sa morphologie et de son contexte syntaxique, une partie de ses attributs grammaticaux.

7.1. Les solutions morphologiques potentielles

Nous avons vu précédemment que pour toute forme inconnue ou incomplète rencontrée en cours d'analyse, le processus de catégorisation automatique permet de lui attribuer un ensemble solution S (cf. V.6.5). Cet ensemble contient les solutions morphologiques (catégories grammaticales et valeurs des variables grammaticales) attribuables à cette forme, d'après sa morphologie (caractères constituant sa terminaison). Ces solutions potentielles sont stockées et numérotées dans une table des formes inconnues, de manière à pouvoir les retrouver ultérieurement dans la même portion de texte sans relancer le processus de catégorisation automatique. Lors de cette rencontre, il n'est pas nécessaire de réactiver le processus de catégorisation automatique, l'ensemble solution S pour cette nouvelle occurrence, est reconstitué directement à partir de la table des formes inconnues. On rappelle que S est un sur-ensemble de l'ensemble solution noté SM, déterminable hors contexte dans le modèle linguistique (cf. IV.3.2.2.).

7.2. Validation des solutions morphologiques potentielles

Le filtrage syntaxique, par l'exploitation du proche contexte positionnel de la forme analysée (formes précédente et suivante), permet d'affiner l'ensemble S produit par le processus de catégorisation automatique en un ensemble solution éventuellement réduit, noté S_{red}. A l'issue de ce filtrage, les solutions morphologiques de l'ensemble S_{red} d'une forme inconnue, sont considérées comme les solutions potentielles de cette forme. Si S_{red} ne contient qu'un élément, on dira que cette solution est validée car elle constitue l'interprétation unique de la forme concernée pour le contexte rencontré. Cette solution est marquée "validée" parmi les solutions morphologiques consignées dans la table, par l'attribution d'un numéro d'unité lexicale.

Les solutions morphologiques non validées de S sont conservées dans la table des formes inconnues jusqu'au moment de l'acquisition de la forme. Elles peuvent donc à leur tour être validées lors de la rencontre dans le texte d'une nouvelle occurrence de cette forme si le contexte rencontré est différent. Ainsi rien n'interdit de reconnaître dans un même texte plusieurs solutions pour une forme homographe.

Une solution morphologique potentielle peut également être validée par l'application d'un schéma de résolution d'ambiguïté, si elle constitue la seule interprétation possible à l'issue de cette application.

7.3. Acquisition du nouveau vocabulaire

A la fin de l'analyse d'un texte, ou d'une portion de texte, si des formes inconnues ont été rencontrées se pose le problème de leur acquisition. Cette acquisition est nécessaire pour les éventuelles reconnaissances ultérieures de ces formes, dans la mesure où l'on veut pouvoir les identifier.

7.3.1. Acquisition automatique

L'acquisition automatique ne peut être envisagée que pour les formes admettant des solutions morphologiques validées. Elle ne sera effectuée que si de plus, on peut déduire de la morphologie de cette forme le représentant de

l'unité lexicale et l'ensemble des racines et modèles morphologiques correspondants. Nous allons considérer deux exemples permettant d'illustrer cette démarche.

Exemple 1 :

Dans l'exemple d'exécution de l'analyseur que nous avons développé précédemment (cf. 2.2.) nous avons rencontré la forme inconnue "*signalisation*" pour laquelle deux solutions potentielles ont été déterminées :

sm_1 = (SUBC MAS,FEM SIN)

et sm_2 = (ADJQ MAS,FEM SIN)

la solution sm_2 se trouvant sur un chemin pendant à l'issue de l'application des schémas de résolution d'ambiguïté est éliminée, ce qui provoque la validation de sm_1.

Tout mot inconnu ayant un comportement grammatical régulier (cette propriété résultant de l'initialisation du dictionnaire d'analyse cf. 3.1.2.), on peut conclure que le représentant d'unité lexicale de cette forme est "*signalisation*", que de plus il constitue l'unique racine de cette unité et que le modèle associé est celui permettant la construction du pluriel régulier des substantifs. Cette forme peut donc être acquise automatiquement, et la même racine permettra de reconnaître aussi bien la forme singulier que la forme pluriel.

Remarque :

Dans cet exemple une imprécision concernant le genre de la forme demeure, puisque le contexte ne permet pas de trancher.

Exemple 2 :

Considérons que la forme inconnue "*lambrissait*" admette comme solution valide :

sm = (VBCJ 3ème SIN)

tout verbe inconnu étant forcément du 1er ou du 2ème groupe de conjugaison, il faut trancher pour pouvoir déduire l'ensemble des racines permettant d'engendrer toutes les formes du verbe. Or dans ce cas c'est impossible, puisque rien ne permet de choisir le verbe du 1er groupe "*lambrisser*" (cas de la désinence *ais*) par rapport à un verbe du 2ème groupe qui serait "*lambrir*" (cas de la désinence *issais*).

Remarque :

La forme "*lambrisse*" permettrait par contre de retrouver la bonne solution, la désinence "e" étant caractéristique du 1er groupe.

Deux cas peuvent donc se produire, correspondant aux deux exemples que nous venons de voir :

- le premier autorise après la détermination du représentant d'unité lexicale, des différentes racines et modèles morphologiques associés, pour la solution morphologique concernée, une acquisition automatique de la nouvelle forme. Cette acquisition s'effectue en fin d'analyse. Toutes les occurrences reconnues dans le texte sont identifiées par le numéro d'unité lexicale ayant servi à marquer la validation de la solution.

- le deuxième ne permet pas cette acquisition. La déduction automatique des renseignements nécessaires n'ayant pu s'effectuer de manière univoque. Une solution demeure néanmoins, à savoir l'acquisition différée sous le contrôle de l'utilisateur.

Pour réaliser, dans le premier cas, cette acquisition automatique, il est obligatoire de disposer de certaines connaissances afin de déduire automatiquement les informations linguistiques nécessaires. Ces connaissances sont disponibles sous la forme de modèles d'acquisition qui sont prédéfinis. Le problème se ramène à attribuer à une solution morphologique validée le bon modèle. Cette attribution est gouvernée par les informations grammaticales de la solution validée et par la morphologie de la forme.

Un modèle d'acquisition est constitué d'un ensemble de racines modèles et des modèles morphologiques associés. Dans le cas des substantifs singuliers, quatre modèles sont prédéfinis. On rappelle que, lors de l'initialisation du dictionnaire, pour la catégorie des substantifs, toutes les formes admettant un pluriel irrégulier ont été consignées.

1) les formes dont la terminaison est "*s*", "*x*", ou "*z*", qui ne changent pas au pluriel;

2) les formes se terminant par "*al*" qui font "aux au pluriel;

3) les formes se terminant par "*au*" ou "*eu*" qui prennent un '*x*' au pluriel;

4) les formes ayant une autre terminaison, qui prennent un "*s*" au pluriel;

Exemple :

Pour la forme "*signalisation*", la catégorie grammaticale et la valeur de la variable nombre permettent de rechercher le modèle d'acquisition parmi les quatre définis pour les substantifs singuliers en fonction de la morphologie terminale de la forme. La

terminaison "*n*" permet dans cet exemple de choisir le modèle n° 4 qui correspond aux formes prenant un "*s*" au pluriel.

7.3.2. Acquisition différée contrôlée

Les solutions morphologiques validées qui ne peuvent pas être acquises automatiquement sont néanmoins stockées avec les informations linguistiques déterminées de manière à ce que l'utilisateur puisse procéder par la suite à une acquisition contrôlée. Il pourra ainsi trancher dans le cas d'informations ambiguës.

On peut ainsi résoudre aisément le problème posé dans l'exemple 2 de la section précédente, en proposant les deux solutions. L'utilisateur n'a pas à intervenir dans la détermination des renseignements linguistiques, ce qui nécessiterait une bonne connaissance du modèle utilisé. Il se contente de choisir une interprétation.

7.4. Conclusion sur l'enrichissement automatique

Ce processus d'enrichissement automatique (complété par une possibilité d'acquisition différée contrôlée) permet d'aborder avec cette analyse des textes dont le vocabulaire est "ouvert". En effet, le processus d'analyse, grâce à un calcul automatique des solutions potentielles, n'est pas bloqué par la rencontre de mots inconnus. Il permet de plus d'affranchir l'utilisateur de la connaissance du modèle linguistique utilisé.

L'initialisation du dictionnaire d'analyse avec les irrégularités grammaticales du français permet un traitement simple des formes incomplètes ou inconnues; même si la détermination de la bonne solution n'est pas toujours possible, en particulier en raison d'un travail encore incomplet au niveau des listes de terminaison des classes ouvertes.

La constitution d'un dictionnaire d'analyse conséquent se réalise à l'aide de ce processus de manière progressive tout en permettant une analyse satisfaisante des textes abordés.

8. Conclusion

L'analyseur morphosyntaxique de surface que nous venons de présenter répond à notre objectif principal de reconnaissance, au sein des syntagmes nominaux, des Goupes Conceptuels. Son caractère partiel tant au niveau de la désambiguïsation grammaticale que des traitements syntaxiques, est compensé sur le pan qualitatif par les possibilités de ciblage, notamment au niveau des résolutions des ambiguïtés grammaticales, et sur le plan quantitatif par la relative simplicité des traitements mis en oeuvre.

Les ensembles de données initiaux nécessitent un collationnement important qui requiert une attention soutenue et une bonne connaissance du modèle utilisé. Mais ils permettent un fonctionnement immédiat de l'analyseur et peuvent s'enrichir progressivement en fonction des corpus traités. Enfin, leur caractère général doit pouvoir s'adapter à toute application particulière.

CHAPITRE VI

PLAN DU CHAPITRE VI

REALISATION ET EXPERIMENTATION

Réalisation et Expérimentation

1. Introduction

Ce chapitre comporte des indications relatives à l'implantation de l'analyseur de surface et du processus d'extraction des Groupes Conceptuels, réalisée sur les machines successives du Laboratoire de Génie Informatique de Grenoble : (VAX 11/780 et GOULD UTX/32) fonctionnant sous UNIX.

Nous présentons les résultats obtenus pour chacun des niveaux d'analyse développés, appliqués à des textes issus de plusieurs corpus, à partir des mêmes ensembles de données linguistiques.

Ce sont ces résultats expérimentaux qui permettent la validation de la méthode que nous proposons, même si les analyseurs réalisés ne constituent que des prototypes.

2. Réalisation

Nous avons réalisé quatre modules qui correspondent à quatre niveaux d'analyse différents. Chacun de ces modules intègre les niveaux d'analyse inférieurs. Ce sont dans l'ordre :

- un analyseur morphologique,
- un module de filtrage syntaxique,
- un analyseur de surface,
- un processus d'extraction des Groupes Conceptuels.

Ce système est complété par un certain nombre d'utilitaires, qui sont indispensables pour la gestion des ensembles de données que nécessite le fonctionnement de ces analyseurs. Ce sont :

- pour le dictionnaire :
  - un utilitaire de construction qui permet la génération des structures de données et l'initialisation; ce même utilitaire permet un enrichissement en mode "batch",
  - un utilitaire d'enrichissement qui permet de compléter interactivement ce dictionnaire,
  - un utilitaire de réorganisation qui permet d'optimiser les accès en privilégiant les racines les plus fréquentes,
  - des utilitaires d'édition et de génération des formes contenues dans le dictionnaire; ces utilitaires sont utiles pour vérifier l'exactitude du vocabulaire.

- pour la matrice de précédence :
  - un utilitaire d'enrichissement et de consultation interactif,
  - un utilitaire d'édition;

- pour les schémas de résolution d'ambiguïté :
  - un utilitaire de saisie et de contrôle des schémas.

Ces programmes sont écrits en langage "Pascal". Ils utilisent des primitives écrites en langage "C" pour gérer les accès aux fichiers du dictionnaire. L'ensemble des modules d'analyse représente un peu plus de 15.000 lignes de code, le plus gros faisant environ 4500 lignes. L'ensemble des utilitaires comporte un peu plus de 5.000 lignes.

L'exécutable réalisant l'extraction des G.C. occupe 220 K., et l'ensemble des fichiers de données nécessaires à son fonctionnement (dictionnaire, matrice, catalogue des schémas, etc ...) environ 600 K.

Le dictionnaire le plus complet, appelé par la suite "dictionnaire enrichi" comporte quelques 57000 formes pour environ 3250 unités lexicales, ce qui a nécessité environ 8400 racines. Dans sa version minimale (classes fermées, irrégularités, et verbes les plus fréquents) il comporte quelques 52800 formes pour environ 2650 unité lexicales ce qui représente quelques 7600 racines.

2.1. L'analyseur de surface

C'est le module d'analyse morphologique qui contient les primitives "C" d'accès aux fichiers du dictionnaire.

Les données linguistiques gérables par les utilitaires présentés dans la section précédente sont maintenues dans des fichiers de type "texte". Ce sont pour l'analyseur de surface :

- les modèles morphologiques (suites de numéros de table de désinence, qui elles sont figées et compilées en même temps que les sources des programmes).
- la matrice de précédence
- le catalogue des schémas de résolution d'ambiguïté.

En début de session, ces données linguistiques sont lues à partir de ces fichiers "texte".

Une fenêtre permettant de mémoriser une suite de formes interprétées et les transitions associées, est créée. La taille de cette fenêtre est un des paramètres de l'analyseur (la valeur définie dans le prototype est 15). C'est ce mécanisme de fenêtre qui permet de stocker les informations constituant le réseau (du moins, la portion "traitable" du réseau).

Le texte du document à analyser est lu forme par forme, par une procédure spécialisée qui permet de traiter directement (i.e. déterminer un ensemble solution) les cas particuliers que sont les nombres, les sigles, les signes de ponctuation, etc... Pour chaque mot rencontré, on recherche dans le dictionnaire un radical permettant de l'analyser (segmentation racine + désinence). En cas de succès, l'ensemble solution, constitué de triplets *"représentant d'unité lexicale, catégorie grammaticale, ensemble de valeurs*

grammaticales" est directement dérivé des renseignements linguistiques stockés dans le dictionnaire. En cas d'échec, on active la procédure de catégorisation automatique qui permet de déduire de la morphologie du mot inconnu, un ensemble solution hors contexte (cf. V.6).

Dès que l'on dispose d'un ensemble solution pour la forme "courante", on peut calculer la transition associée à la forme "précédente" (mémorisée dans le premier élément de la fenêtre). On décale ensuite la fenêtre d'une forme. L'ensemble solution déterminé est mémorisé avec la chaîne de caractères constituant la forme courante dans le premier élément de la fenêtre. La transition associée ne peut être définie avant que la forme suivante ne soit analysée. La figure 1 illustre ce processus séquentiel :

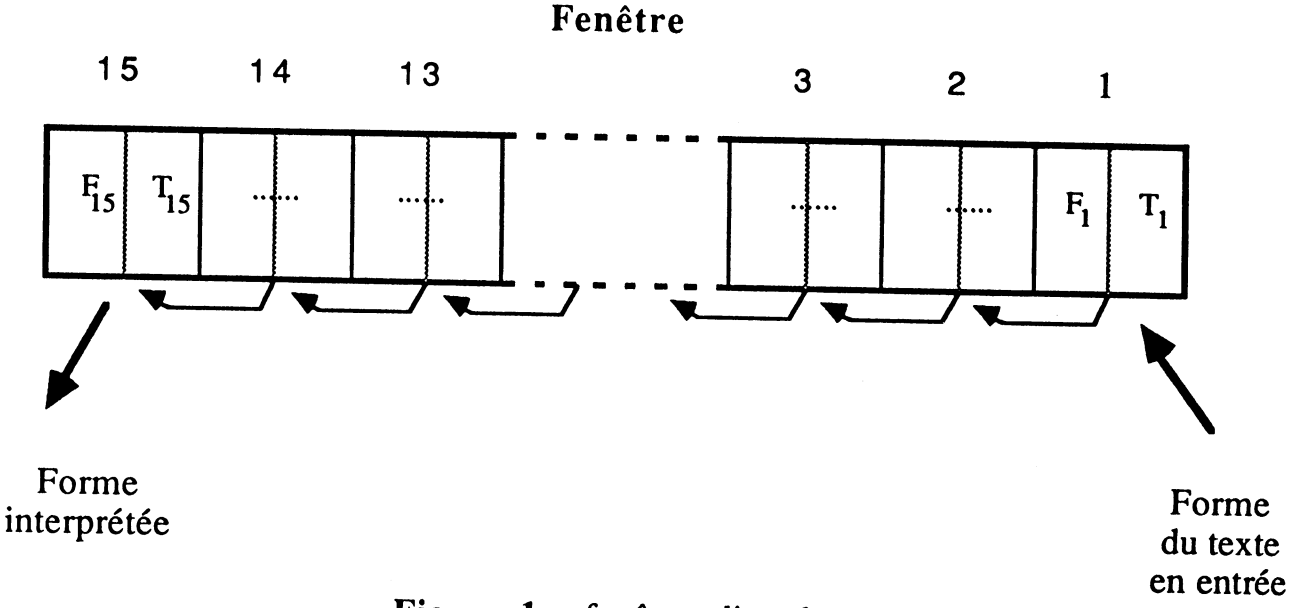


Figure 1 : *fenêtre d'analyse*

Les processus de filtrage syntaxique sont appliqués sur la transition d'un élément de la fenêtre, dont le rang r_f est un paramètre de l'analyse (ce rang correspond au premier élément dans le prototype).

De même, les processus de validation et d'application des schémas de résolution d'ambiguïté, et de nettoyage du réseau sont appliqués à partir d'un élément de la fenêtre dont les rang r_s et r_n sont également paramétrables.

Si l'on désigne par CS le catalogue des schémas de résolution d'ambiguïté, CSA le catalogue des schémas activés, CSV le catalogue des schémas validés, et que l'on donne le rang de l'élément de la fenêtre sur lequel sont appliqués les

traitements entre crochets [r], on peut représenter cette analyse par l'algorithme suivant :

```
Tanque il reste une forme à analyser Faire
    lire une forme
    analyser la forme courante
    créer la transition [1]
    filtrer [rf]      /* en utilisant la matrice de précédence*/
    décaler la fenêtre
    ranger la forme courante dans la fenêtre [1]
    poursuivre la reconnaissance des schémas activés [rs]  /* dans CSA */
    essayer d'activer de nouveaux schémas [rs]   /* dans CS */
    appliquer les schémas validés [rs]    /* dans CSV */
    nettoyer le réseau [rn]
fin tanque
```

En fonction de la taille des schémas de résolution d'ambiguïté définis dans le catalogue, on peut "jouer" sur la taille de la fenêtre et sur les rangs d'application des traitements (r_f ,r_s et r_n).

2.2. L'extraction des Groupes Conceptuels

Le processus d'extraction des G.C. intègre l'analyseur de surface. En plus des données linguistiques nécessaire à l'analyse, en début de session, les données constituant la table de transition du transducteur sont lues à partir d'un fichier "texte". Ainsi, un changement de la syntaxe des structures à reconnaître ne nécessite qu'une modification des données de ce fichier.

L'entrée du transducteur est constitué par la forme interprétée "sortant" de la fenêtre à chaque décalage. Les tests de linéarité du réseau (nécessaire pour les portions traitées par le transducteur) sont réalisés sur la transition associée à la forme en sortie de la fenêtre.

La figure 2 ci-après, permet de représenter l'enchaînement des traitements réalisés par l'analyseur de surface et par le processus d'extraction des Groupes Conceptuels.

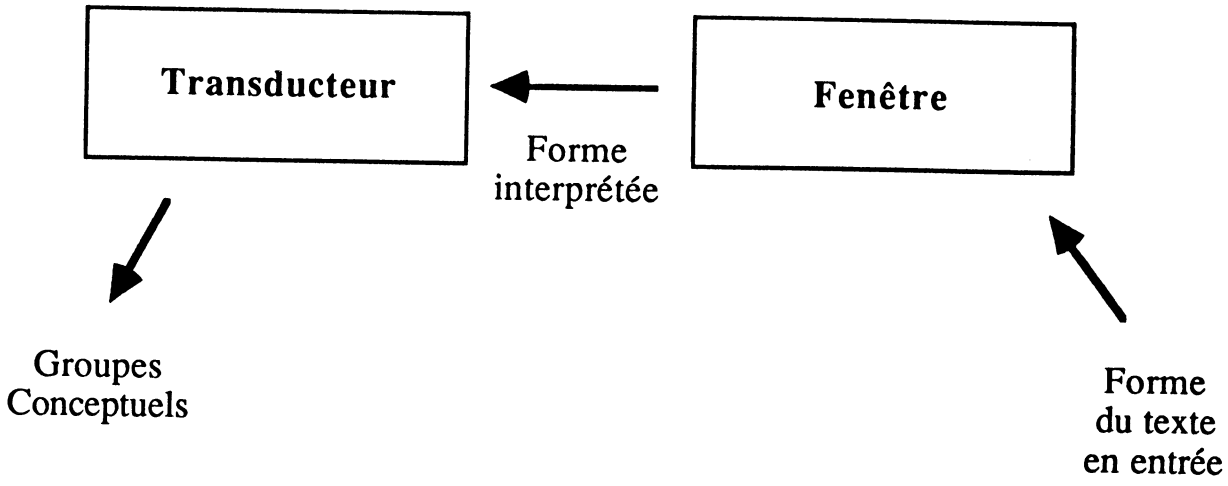


Figure 2 : *extraction des Groupes Conceptuels*

A la fin du texte d'entrée, le décalage s'effectue jusqu'à ce que la fenêtre soit vidée.

3. Expérimentation

Nous avons réalisé une expérimentation sur des textes issus de trois corpus différents. Avant de présenter ces corpus, il convient de souligner la difficulté d'évaluation comparative d'une méthode, sans un corpus d'expérimentation étalonné.

Le premier corpus est constitué par les NEF (manuel de normalisation pour les autocommutateurs, utilisé par les ingénieurs du CNET). Ce corpus a constitué la base d'expérimentation du prototype IOTA (cf. I.5). Il s'agit d'un corpus technique dont le vocabulaire et la syntaxe sont relativement pauvres. La phrase exemple qui a illustré le chapitre V est tirée de cette documentation. Nous avons retenus deux chapitres de ce manuel, le plus long et le plus court.

Nous avons appliqué nos méthodes sur un texte extrait d'un deuxième corpus, utilisé dans le cadre du projet RIME [BERR 88] et [NIE 90], qui est constitué de comptes rendus médicaux. Ce sont des documents relativement courts, rédigés par des spécialistes (médecins), qui décrivent l'examen subi par

un malade et qui en tirent des constatations. Le vocabulaire de spécialité utilisé dans ces textes, bien que technique, est totalement différent de celui rencontré dans les documents des NEF. Le style de rédaction, très concis et précis, renforce l'utilisation de ce vocabulaire de spécialité.

Chacun de ces deux corpus nécessiterait un apprentissage particulier du vocabulaire, qui permettrait d'en cerner rapidement la quasi-totalité. Ils mettent bien en évidence les "barrières" cloisonnant les langues de spécialité.

Enfin nous avons choisi d'expérimenter cet analyseur sur l'introduction de ce mémoire.

Nous désignerons par la suite, par "ch4" et "ch13" les deux chapitres extraits du corpus des NEF, par "crmed" un texte extrait du corpus des comptes rendus médicaux, et par "intro" un texte ayant servi à l'introduction de ce mémoire. La table ci-après permet d'apprécier les tailles respectives de ces textes :

textes	nombre de formes	nombre de signes de ponctuation
ch4	922	113
ch13	7111	875
crmed	197	34
intro	1387	142

On pourra trouver en annexe des extraits des textes analysés provenant de ces corpus (cf. annexe 6).

Cette expérimentation a été réalisée, pour chacun des textes de ces corpus, avec deux versions réorganisées du dictionnaire (ce qui procure un gain d'environ 10% du temps d'exécution) :

- la première est constituée par le dictionnaire "minimal", construit à partir des éléments des classes fermées, des irrégularités, et des éléments les plus fréquents des classes ouvertes (cf V.6.1).

- la deuxième dit "enrichi" correspond à un enrichissement du dictionnaire "minimal", réalisé à partir d'un texte des NEF, différent de ceux servant à l'expérimentation, qui a été analysé manuellement. Nous avons volontairement réduit cet enrichissement de manière à rester dans un contexte d'appréhension d'univers textuels ouverts, c'est-à-dire dont le vocabulaire connu est toujours incomplet. Le texte comporte 3882 formes et 545 signes de ponctuation.

Nous donnons sur le graphique suivant les taux de linéarisation observés pour ce texte, en fonction du taux de reconnaissance du vocabulaire, pour les modules de filtrage syntaxique (avec nettoyage du réseau résultat), et d'analyse de surface. On peut constater que la connaissance totale du vocabulaire ne permet une linéarisation que de 66,5% avec le filtrage, et 78% avec l'application des schémas de résolution d'ambiguïté.

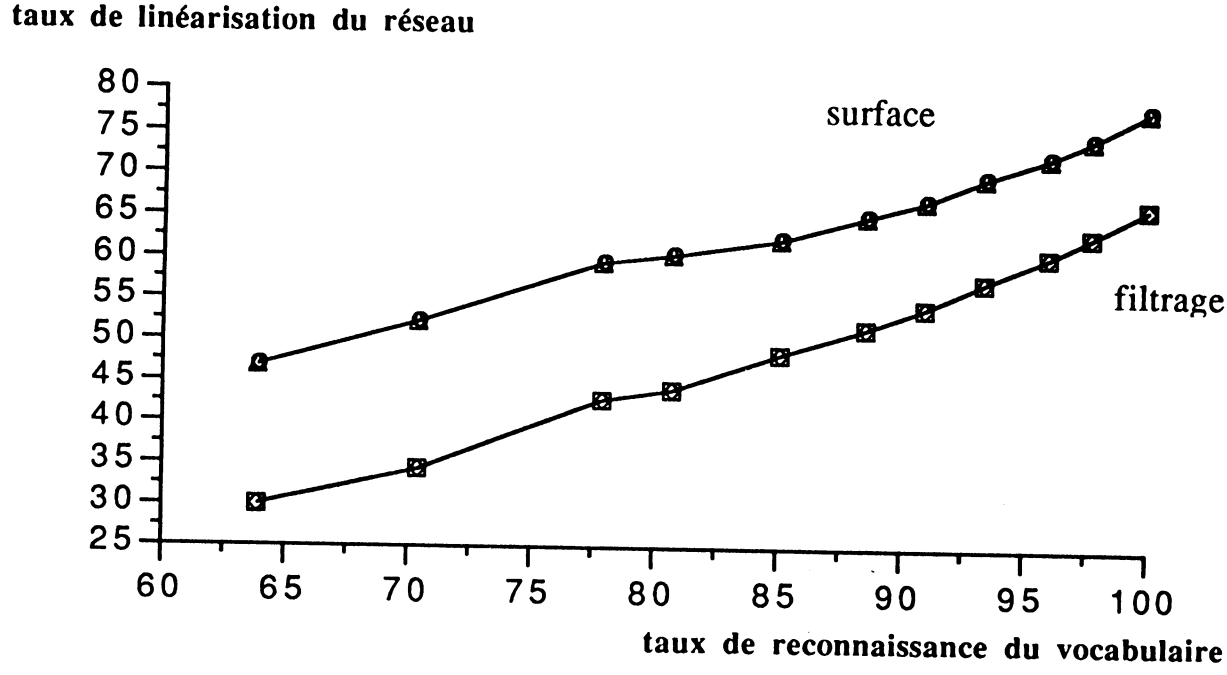


Figure 3 : *influence du taux de reconnaissance du vocabulaire sur la linéarisation du réseau*

Ces résultats ont été obtenus avec d'une part, un catalogue comprenant 29 schémas de résolution d'ambiguïté, définis pour la plupart à partir d'ambiguïtés classiques, et d'autre part une matrice de précédence considérée comme complète (les successions non définies sont considérées comme impossibles).

Ce sont ces mêmes ensembles de données qui ont été utilisés pour les expérimentations dont nous présentons maintenant les résultats par niveau d'analyse.

3.1. Morphologie

Nous donnons dans le tableau 1 ci-après, les résultats obtenus par l'analyseur morphologique sur les textes des différents corpus. Les colonnes "nbfa", "nbfi", "tr" et "temps" représentent respectivement pour 1000 formes analysées (la ponctuation n'étant pas comptabilisée) : le nombre de formes inconnues, le taux de reconnaissance, et le temps d'exécution (cpu) en secondes obtenu sur la machine GOULD (machine bi-processeur 2 x 7 Mips). La ligne "moy" donne les valeurs moyennes (calculées pour 1000 formes) obtenues sur l'ensemble des textes analysés, la ligne "moy pond" correspond à une moyenne calculée pour 1000 formes de chaque corpus.

	dictionnaire "minimal"			dictionnaire "enrichi"		
textes	**nbfi**	**tr**	**temps**	**nbfi**	**tr**	**temps**
ch4	340	66%	24.5 s	190	81%	24.7 s
ch13	338	66.2%	23.4 s	204	79.6%	24.5 s
crmed	523	47.7%	22.3 s	492	50.8%	23.5 s
intro	376	62.4%	23.5 s	218	78.2%	24.3 s
moy	394	60.6%	23.4 s	276	72.4%	24.3 s
moy pond	413	58.7%	23.2 s	302	69.8%	24.1 s

Tableau 1 : *résultats de l'analyse morphologique.*

On voit apparaître clairement sur ce tableau, les caractéristiques des vocabulaires des différents corpus : le dictionnaire "minimal" permet de reconnaître les deux-tiers du vocabulaire utilisé dans les NEF et seulement un peu moins de la moitié de celui utilisé dans les comptes rendus médicaux. L'évolution du taux de reconnaissance avec l'utilisation du dictionnaire "enrichi" (à partir d'un texte des NEF) démontre qu'une partie importante du vocabulaire appris, n'est pas spécialisé. En effet pour le texte de notre introduction, l'évolution de ce taux est sensiblement équivalente à celle observée pour les deux chapitres des NEF analysés.

Les petites différences de temps d'exécution observées avec les deux versions du dictionnaire, s'expliquent par le fait que les écritures permettant de constituer la trace d'exécution sont plus nombreuses lorsque le nombre de formes reconnues est plus important.

Ce temps "cpu" est la somme du temps "utilisateur" correspondant au temps de calcul du programme, et du temps "système" décomptant en particulier les entrées-sorties, fournis par la commande "time" sous UNIX.

3.2. Filtrage positionnel et grammatical

Nous donnons dans les tableau 2 et 3 ci-après, les résultats obtenus par le filtrage positionnel et grammatical pour les textes sélectionnés. Le tableau 2 comporte les résultats "bruts" du filtrage, sans élimination des chemins pendants. Le tableau 3 tient compte de cette élimination. Les colonnes "nbfp", "nbap", "nbai", "ts" et "temps" représentent respectivement pour 1000 formes analysées (la ponctuation n'étant pas comptabilisée) : le nombre de formes et de ponctuations, le nombre d'arcs potentiels, le nombre d'arcs invalidés, le taux de simplification, et le temps d'exécution (cpu) en secondes obtenu sur la machine GOULD. La ligne "moy" donne les valeurs moyennes (calculées pour 1000 formes) obtenues sur l'ensemble des textes analysés, la ligne "moy pond" correspond à une moyenne calculée pour 1000 formes de chaque corpus.

	dictionnaire "minimal"					dictionnaire "enrichi"				
textes	**nbfp**	**nbap**	**nbai**	**ts**	**temps**	**nbfp**	**nbap**	**nbai**	**ts**	**temps**
ch4	1123	3716	617	16.6%	32.4 s	1123	3012	452	15%	29.7 s
ch13	1123	3527	589	16.7%	32.1 s	1123	2928	410	14%	30 s
crmed	1173	4822	685	14.2%	34 s	1173	4721	685	14.5%	33 s
intro	1102	4264	924	21.7%	33.4 s	1102	3524	714	20.3%	31.5 s
moy	1130	4082	704	17.2%	33 s	1130	3546	565	15.9%	31 s
moy pond	1133	4236	737	17.4%	33.2 s	1133	3738	610	16.3%	31.5 s

Tableau 2 : *résultats du filtrage positionnel et grammatical.*

On peut remarquer sur le tableau 2 que la complexité du réseau potentiel découlant de la morphologie, donnée par le nombre d'arcs potentiels "nbap" décroît lorsque le taux de reconnaissance du vocabulaire "tr" (tableau 1) augmente. Cela provient du fait que les formes inconnues sont interprétées par le processus de catégorisation automatique qui fournit un ensemble solution hors-contexte, qui est un sur-ensemble de l'ensemble solution "exact", et qui donc contenant plus de solutions morphologiques génère plus d'arcs.

Les taux de simplification "ts" du tableau 3 correspondent à la simplification réelle résultant du filtrage : en moyenne 25% de la complexité du réseau, même si pour faciliter la détermination des parties gauches des schémas (cf. V.5.5.1), ce nettoyage n'intervient qu'après le processus d'application des schémas de résolution d'ambiguïté.

	dictionnaire "minimal"					dictionnaire "enrichi"				
textes	**nbfp**	**nbap**	**nbai**	**ts**	**temps**	**nbfp**	**nbap**	**nbai**	**ts**	**temps**
ch4	1123	3716	905	24.4%	28.5 s	1123	3012	701	23.3%	29.8 s
ch13	1123	3527	862	24.4%	26.9 s	1123	2928	664	22.7%	29.7 s
crmed	1173	4822	974	20.2%	28.3 s	1173	4721	1036	21.9%	32.4 s
intro	1102	4264	1360	31.9%	27 s	1102	3524	1126	32%	32 s
moy	1130	4082	1025	25.1%	27.7 s	1130	3546	882	24.9%	31 s
moy pond	1133	4236	1072	25.3%	27.7 s	1133	3738	948	25.4%	31.4 s

Tableau 3 : *résultats du filtrage positionnel et grammatical avec suppression des chemins pendants.*

Les temps d'exécutions observés dans le tableau 3 sont inférieurs à ceux du tableau 2, alors qu'un traitement a été ajouté, du fait que le fichier texte contenant la trace d'exécution est plus petit (moins d'arcs persistants), ce qui correspond à moins d'ordres d'écriture.

3.3. Analyse de surface

Nous donnons dans le tableau 4 ci-après les résultats obtenus par l'analyseur de surface (après application des schémas de résolution d'ambiguïté et nettoyage du réseau), sur les textes des différents corpus. Les colonnes "nbarf", "nbae", "ts", "linéar." et "temps" représentent respectivement pour 1000 formes analysées (la ponctuation n'étant pas comptabilisée) : le nombre d'arcs résultant du filtrage, le nombre d'arcs éliminés par l'application des schémas de résolution d'ambiguïté et par le nettoyage du réseau, le taux de simplification par rapport au réseau filtré, la linéarité du réseau résultat en pourcentage, et le temps d'exécution (cpu) en secondes obtenu sur la machine GOULD. La ligne "moy" donne les valeurs moyennes (calculées pour 1000 formes) obtenues sur l'ensemble des textes analysés, la ligne "moy pond" correspond à une moyenne calculée pour 1000 formes de chaque corpus.

Nous retrouvons sur ce tableau 4, pour le taux de linéarisation du réseau, le même type de résultats que celui obtenu lors de l'enrichissement du dictionnaire (cf. figure 3), à savoir : que le taux de linéarisation est d'autant meilleur que le taux de reconnaissance du vocabulaire est élevé. Cette constatation relativise les résultats obtenus, puisque les versions du dictionnaire dont on dispose sont loin de permettre des taux très élevés.

	dictionnaire "minimal"					dictionnaire "enrichi"				
textes	**nbarf**	**nbae**	**ts**	**linéar.**	**temps**	**nbarf**	**nbae**	**ts**	**linéar.**	**temps**
ch4	3099	1023	33%	56.8%	32.6 s	2560	803	31.4%	68.7%	30.4 s
ch13	2938	859	29.2%	50.9%	32.1 s	2518	653	25.9%	60.3%	30 s
crmed	4137	817	19.7%	33%	34 s	4036	817	20.2%	34.8%	33.5 s
intro	3340	955	28.6%	43.4%	32.4 s	2810	750	26.7%	52.7%	31.5 s
moy	3378	914	27%	46%	32.8 s	2981	756	25.4%	54.1%	31.4 s
moy pond	3499	904	25.8%	43.4%	32.9 s	3128	742	23.7%	50.7%	31.7 s

Tableau 4 : *résultats de l'analyseur de surface.*

Le tableau 5 ci-après contient le nombre d'occurrences de schémas de résolution d'ambiguïté appliqués "nbsa", et ce même nombre d'occurences évalué pour 1000 mots "nbsapm".

	dictionnaire "minimal"		dictionnaire "enrichi"	
textes	**nbsa**	**nbsapm**	**nbsa**	**nbsapm**
ch4	246	267	185	201
ch13	1576	222	1134	159
crmed	35	178	33	168
intro	265	191	178	128
moy	-	214.5	-	164
moy pond	-	204.5	-	159

Tableau 5 : *schémas appliqués*

Il est important de préciser que ces résultats ont été obtenus avec un nombre de schémas de résolution d'ambiguïté relativement réduit. L'apprentissage poussé permettrait probablement d'en acquérir beaucoup plus. Notre objectif était ici de vérifier la faisabilité de cette démarche.

Néanmoins, les résultats obtenus tant sur le plan qualitatif que quantitatif, avec le processus d'extraction des Groupes Conceptuels, à partir de ces analyses "incomplètes", corroborent nos hypothèses et valident cette méthode.

3.4. Extraction des Groupes Conceptuels

L'expérimentation a été réalisée pour l'extraction des Groupes Conceptuels à partir d'un automate défini par la table de transition donné au chapitre III (cf. III.3). Il s'agit donc d'une version incomplète qui ne prend pas en compte l'ensemble de la syntaxe cible définie pour les G.C.

Nous donnons dans le tableau 6 ci-après les résultats obtenus par ce processus pour les textes sélectionnés. Les colonnes "nbgc", "cmax", "cmoy" et "temps" représentent respectivement le nombre de G.C. extraits, la complexité maximale des G.C. extraits donnée en nombre de formes, la complexité moyenne et le temps d'exécution calculé en secondes obtenu sur la machine GOULD.

	dictionnaire "minimal"				dictionnaire "enrichi"			
textes	**nbgc**	**cmax**	**cmoy**	**temps**	**nbgc**	**cmax**	**cmoy**	**temps**
ch4	84	5	1.57	27.1 s	120	5	1.74	29 s
ch13	711	5	1.38	206.5 s	1008	5	1.49	215.7 s
crmed	12	2	1.25	6.2 s	18	2	1.17	6.4 s
intro	132	4	1.39	39.5 s	182	5	1.48	43.4 s

Tableau 6 : *résultats du processus d'extraction des G.C.*

Voici quelques groupes nominaux extraits de ces textes débouchant sur des Groupes Conceptuels (les mots soulignés sont des mots inconnus qui ont été interprétés correctement par l'analyseur) :

"signaux de ligne du code de signalisation pour l'accès" (ch4)

↓

signal ***de*** ligne ***de*** **det** code ***de*** signalisation ***pour*** **det** accès

Ce premier groupe est intéressant d'une part parce qu'il provient de la phrase exemple illustrant le chapitre V, et d'autre par parce qu'il est incomplet, le groupe nominal complet étant :

"signaux de ligne du code de signalisation pour l'accès à l'interurbain manuel"

l'ambiguïté persistante pour les formes *"interurbain"* et *"manuel"* entre les interprétations SUBC et ADJQ (substantif commun et adjectif qualificatif), empêche l'extraction complète.

"permettre à un opérateur d'un centre de commutation de localiser" (ch13)

↓

permettre ***à*** **det** opérateur ***de*** **det** centre ***de*** commutation ***de*** localiser

Ce G.C. a la particularité de faire intervenir deux verbes à l'infinitif. L'absence du complément suivant la forme "localiser" empêche l'appréhension de la signification complète associée à ce groupe. Les extensions prévues au troisième chapitre, pour la prise en compte des syntagmes verbaux permettraient de réduire cet effet (cf. III.2.5).

"aspect d'adénopathie" (crmed)

↓

aspect ***de*** adénopathie

Il est intéressant de constater que ce G.C est constitué avec deux formes "apprises" lors de l'analyse et correctement interprétées par le processus de traitement des mots inconnus.

"fonction d'indexation du système de recherche d'informations" (intro)

↓

fonction ***de*** indexation ***de*** **det** système ***de*** recherche ***de*** information

Ce dernier groupe fait en réalité partie d'un groupe adjectival attribut qui est :

"*réalisée par la fonction d'indexation du système de recherche d'informations*"

la prochaine version du transducteur intégrera ce type de reconnaissance prévu dans la syntaxe cible des G.C..

Le tableau 7, ci-après donne les mêmes informations que le tableau 6, calculées pour 1000 formes analysées, ce qui permet de comparer les résultats

obtenus pour chaque corpus. La ligne "moy" donne les valeurs moyennes (calculées pour 1000 formes) obtenues sur l'ensemble des textes analysés, la ligne "moy pond" correspond à une moyenne calculée pour 1000 formes de chaque corpus.

	dictionnaire "minimal"				dictionnaire "enrichi"			
textes	**nbgc**	**cmax**	**cmoy**	**temps**	**nbgc**	**cmax**	**cmoy**	**temps**
ch4	91	5	1.57	29.4 s	130	5	1.74	31.5 s
ch13	100	5	1.38	29 s	142	5	1.49	30.3 s
crmed	61	2	1.25	31.5 s	91	2	1.17	32.5 s
intro	95	4	1.39	28.5 s	131	5	1.48	31.3 s
moy	87	4	1.4	29.6 s	124	4.25	1.47	31.4 s
moy pond	84	3,7	1.37	29.7 s	119	4	1.42	31.6 s

Tableau 7 : *résultats calculés pour 1000 formes analysées, du processus d'extraction des G.C.*

Le texte crmed, issu du corpus des comptes rendus médicaux, de part son faible taux de reconnaissance du vocabulaire, limite très sensiblement la reconnaissance des G.C.

4. Conclusion

Les résultats que nous venons de présenter permettent de valider la démarche développée en démontrant que les outils linguistiques réalisés constituent un compromis intéressant entre les analyseurs classiques et les outils très spécifiques qui peuvent être développés dans le domaine de la recherche d'informations pour la réalisation de fonctions d'indexation automatique.

Ces résultats ont été obtenus avec des prototypes dont l'ambition est de démontrer la faisabilité de telles approches.

Nous développons actuellement un processus d'extraction de groupes conceptuels que l'on peut qualifier de "tolérant" à certains types d'ambiguïté

persistante : par exemple, en regroupant au moment de cette extraction les catégories SUBC et ADJQ. D'ailleurs, pour de nombreux linguistes cette différenciation n'a pas de raison d'exister en français. Ce type de regroupement permet d'obtenir pour le groupe nominal discuté dans la section précédente :

"signaux de ligne du code de signalisation pour l'accès à l'interurbain manuel"

et de déduire le G.C. suivant :

signal ***de*** ligne ***de***_**det** code ***de*** signalisation ***pour***_**det** accès ***à*** **det** interurbain manuel

CONCLUSION

D'une manière générale, le niveau de structuration des termes d'indexation conditionne l'aspect qualitatif de la "compréhension" du contenu des documents d'un corpus. C'est cette "compréhension" qui permet de représenter le contenu des documents dans un modèle sémantique adéquat. L'analyse linguistique lors de la phase d'indexation permet par la reconnaissance automatique de concepts structurés, à partir du contenu textuel des documents, cette représentation. Les concepts reconnus constituent alors des éléments du modèle sémantique. La fonction de correspondance, établie entre cette représentation du contenu des documents et les thèmes exprimés dans la requête d'un utilisateur, constitue l'essentiel de la problématique des systèmes de recherche d'informations.

La définition d'outils linguistiques pour les SRI nécessite la prise en compte de paramètres spécifiques à ce domaine:

- appréhension d'univers textuels ouverts, et donc capacité à analyser des mots inconnus;

- aptitude à traiter des corpus volumineux (pour la fonction d'indexation), ce qui nécessite des outils d'analyse performants.

Les outils réalisés, qui sont l'analyseur de surface et le processus d'extraction des Groupes Conceptuels, constituent le module d'analyse linguistique de la fonction d'indexation du prototype de système de recherche d'informations IOTA. Cette analyse permet de reconnaître automatiquement, à partir de leur contexte textuel, les concepts structurés susceptibles d'indexer un document. Ces mêmes outils pourront servir de base d'analyse pour les requêtes d'interrogation exprimées en langage naturel.

Tout en satisfaisant les spécificités de la recherche d'informations, nous avons tenu à conserver à cet analyseur, par la définition d'un modèle linguistique proche des modèles classiques, un caractère général qui permet d'envisager une plus vaste utilisation. La classification syntaxique utilisée, notamment de par sa simplicité et son classicisme, doit permettre aisément cette généralité.

Sur le plan méthodologique, si l'analyse morphologique réalisée est dans son ensemble classique, les traitements syntaxiques ne concernent que la structure de surface des portions de texte analysés en combinant un filtrage positionnel et grammatical et un processus de résolution d'ambiguïtés grammaticales ciblées:

- Le filtrage est basé d'une part sur les relations positionnelles (i.e. les restrictions portant sur les possibilités de succession) des catégories grammaticales, et d'autre part sur les contraintes d'accord grammatical entre certaines classes syntaxiques du modèle linguistique utilisé. Il s'agit d'une analyse grammaticale simple.

- Le processus de résolution d'ambiguïtés grammaticales s'appuie sur une modélisation qui permet de représenter une ambiguïté et sa solution (ou du moins une simplification) au sein du contexte de résolution. Avec ce mécanisme il est possible de cibler des ambiguïtés jugées intéressantes en fonction d'objectifs particuliers attribués à l'analyseur: nous nous sommes appliqué pour l'extraction des groupes conceptuels à une désambiguïsation des portions de texte susceptibles de contenir des groupes nominaux.

Enfin, le processus de catégorisation automatique détermine un ensemble solution hors contexte pour les mots inconnus. Cette détermination permet une prise en compte du vocabulaire inconnu sans interruption de l'analyse, et dans certains cas (cf. V.7) son acquisition automatique. Ce processus est essentiel pour l'analyse de textes dont le vocabulaire est ouvert. Nous avons signalé au chapitre précédent (cf. VI.3.4), lors de la présentation de quelques G.C. extraits à partir des textes sélectionnés, les mots inconnus correctement interprétés par l'analyse.

L'expérimentation menée sur des textes issus de corpus différents et dont les résultats sont présentés au chapitre précédent, a permis de valider notre approche. En effet l'analyse de surface mise en oeuvre, malgré un vocabulaire très incomplet et un ensemble réduit de schémas de résolution d'ambiguïté, permet une linéarisation suffisante du réseau résultat pour réaliser une extraction satisfaisante des Groupes Conceptuels. Il serait intéressant d'approfondir cette étude à partir d'un dictionnaire d'analyse beaucoup plus conséquent.

Sur le plan pratique, la représentation des connaissances linguistiques nécessaires au fonctionnement de l'analyse, au moyen d'outils commodes à manipuler, tels que les tables de désinences, la matrice de précédence ou les schémas de résolution d'ambiguïté, a facilité la réalisation de l'analyseur. Les performances observées pour cette analyse proviennent de cette relative simplicité des traitements mis en oeuvre.

Il est intéressant de constater que le prototype d'analyse réalisé, tout en étant pour l'instant encore expérimental, fournit des résultats très significatifs. Les principales améliorations pouvant être apportées, concernent essentiellement les données linguistiques:

- en complétant considérablement le dictionnaire d'analyse; il est généralement admis qu'un dictionnaire intéressant pour une analyse du français doit comporter un minimum d'environ 300.000 formes;

- en étudiant les restrictions portant sur les terminaisons possibles des éléments des classes ouvertes, et en particulier pour les substantifs communs et les adjectifs qualificatifs;

- enfin en enrichissant le catalogue des schémas de résolution d'ambiguïté à partir des ambiguïtés persistant dans les textes analysés.

En ce qui concerne l'extraction des Groupes Conceptuels, nous développons actuellement une version permettant la prise en compte des groupes adjectivaux tels qu'ils sont définis en III.2.2. Ces groupes ne sont pas reconnus par l'automate ayant servi à l'expérimentation détaillée au chapitre VI. L'étape ultérieure sera la reconnaissance des relations de connexion entre les GC. Ces relations (prépositionnelles, auxiliaires et de proximité), telles qu'elles sont définies en III.2.4, permettent de compléter la description sémantique du contenu des documents en donnant une certaine interprétation des syntagmes verbaux. Cette extension doit permettre une meilleure "compréhension" du contenu des documents.

Sur le plan fonctionnel, des extensions sont prévues au niveau des interfaces utilisateur, afin de faciliter notamment la gestion et l'enrichissement des ensembles de données linguistiques du système.

Néanmoins, sur le plan qualitatif, une évaluation sérieuse des résultats obtenus et leur comparaison avec ceux d'autres systèmes, ne pourra être réalisée qu'après l'intégration de ces traitements linguistiques dans un SRI opérationnel. Actuellement le prototype IOTA ne permet pas ce type d'évaluation.

Sur un plan plus prospectif, les développements ultérieurs de ce travail devront porter à notre avis sur les points suivants:

- au niveau analyse, essayer de compléter les traitements effectués par la définition de processus spécialisés pour la prise en compte de certaines particularités linguistiques, telles que les références pronominales simples (par exemple l'emploi de pronoms relatifs dans des expressions syntaxiquement figées), ou certaines tournures stylistiques;

- au niveau des Groupes Conceptuels, étudier les possibilités d'extension de la syntaxe des GC pour la prise en compte d'une part des syntagmes verbaux et de leur interprétation, et d'autre part des connecteurs principaux (conjonctions de coordination "et" et "ou" par exemple);

- une étude plus poussée des relations reconnaissables entre les Groupes Conceptuels doit permettre une aide non négligeable pour toute application contribuant à une constitution automatique de base de connaissances. En particulier, une étude dans ce contexte des rapports exprimés par les prépositions nous parait prometteuse.

Il serait également intéressant, d'expérimenter ce type de traitements linguistiques de surface, pour des langues aux caractéristiques proches du français, i.e avec de fortes contraintes positionnelles et une morphologie riche.

BIBLIOGRAPHIE

REFERENCES BIBLIOGRAPHIQUES

ABRIAL J.R.
Data Semantics.
Data Base Management, Klimbie J.W. et al eds., North Holland, Amsterdam, 1974.

ANDREEWSKY A., FLUHR C.
Apprentissage, analyse automatique du langage, application à la documentation.
Dunod, Document de Linguistique Quantitative n° 21, 1973.

ANDREEWSKY A., DEBILI F. FLUHR C.
Computational learning of semantic lexical relations for the generation and automatical analysis of content.
IFIP, Toronto, août 1977.

ANTONIADIS G.
Elaboration d'un système d'analyse morpho-syntaxique d'une langue naturelle: application à l'indexation automatique.
Thèse de 3ème cycle, Grenoble II, 1984.

BARTHES C., CARPUAT B., FRONTIN J., GLIZE P.
Un système expert en recherche documentaire multibase et multiserveur.
RIAO 85, Grenoble, mars 1985.

BASSANO J.C.
DIALECT un système expert pour la recherche documentaire.
Thèse d'état, Université de Paris-sud, centre d'Orsay, 1986.

BASSANO J.C.
Systèmes experts et systèmes documentaires intelligents.
Les systèmes experts & leurs applications, Avignon, mai 1987.

BELLY N., BORILLO A., VIRBEL J., SIOT-DECAUVILLE N.
Procédures d'analyses sémantiques appliquées à la documentation sémantique.
Gauthier-Villars, 1970.

BERRUT C.

Résolution des ambiguïtés grammaticales. Une première approche dans le cadre d'un analyseur de surface de la langue naturelle.
Rapport de DEA, Grenoble 1985.

BERRUT C., PALMER P.

Solving grammatical ambiguities within a surface syntactical parser of automating indexing.
ACM SIGIR, Pise, septembre 1986.

BERRUT C.

Une méthode d'indexation fondée sur l'analyse sémantique de documents spécialisés. Le prototype RIME et son application à un corpus médical.
Thèse de l'Université Joseph Fourier, Grenoble I, décembre 1988.

BOITET CH., GUILLAUME P., QUEZEL-AMBRUNAZ M.

ARIANE-78: an integrated environment for automated translation and human revision.
COLING 82, Prague, juillet 1982.

BOOKSTEIN A., SWANSON D.R.

A decision theoric foundation for indexing.
Journal of ASIS, janvier-février 1975.

BOSC P., COURANT M., ROBIN S., TRILLING L.

HAVANE: un système de mise en relation automatique de petites annonces.
Rapport INRIA n° 223, juillet 1983

BRUANDET M.F., CHIARAMELLA Y., KERKOUBA D.

Méthodes empiriques de construction de thésaurus: expérimentation.
Bulletin du C.I.D., n° janvier-mars 1983.

BRUANDET M.F.

Partial knowledge model for an information retrieval system.
RIAO 85, Grenoble, mars 1985.

BRUANDET M.F.

Outline of a knowledge base model for an intelligent information retrieval system.
ACM SIGIR, New Orleans, juin 1987.

CHANDIOUX J.
METEO: un système opérationnel pour la traduction automatique des bulletins météorologiques destinés au grand public.
Groupe TAUM, Université de Montréal, 1976.

CHAUCHE J.
Transducteurs & Arborescences. Etudes et réalisations de systèmes appliquées aux grammaires transformationnelles.
Thèse d'Etat, Grenoble 1974.

CHIARAMELLA Y.
Un état de l'art de la recherche en informatique documentaire.
Colloque C.I.D., Paris, octobre 1983.

CHIARAMELLA Y., BRUANDET M.F., KERKOUBA D.
Intégration d'une fonction documentaire dans un atelier logiciel.
Journées CONCERTO, Perros Guirrec, février 1986.

CHIARAMELLA Y., BRUANDET M.F., DEFUDE B., KERKOUBA D.
IOTA: aprototype of an information retrieval system.
ACM SIGIR, Pise, septembre 1986.

CHIARAMELLA Y., DEFUDE B.
IOTA: un prototype de système expert en recherche d'informations.
Les systèmes experts & leurs applications, Avignon, mai 1987.

CHOMSKY N.
Syntactic structures.
Mouton, La Haye, 1957.

CHOMSKY N.
Aspects of theory of syntax.
The M.I.T. Press, 1965.

CHRISTODOULAKIS S.
Multimédia retrieval.
ACM SIGIR, New Orleans, juin 1987.

COLMERAUER A.
Systèmes-Q.
Rapport TAUM 71, Université de Montréal, 1971.

COLMERAUER A.
Les grammaires de métamorphose.
Groupe d'intelligence artificielle, Université d'Aix-Marseille II 1975.

COLMERAUER A.
Total precedence relations.
Journal of the ACM, vol 17, n° 1, janvier 1970.

COLMERAUER A., KANOUI H., VAN CANEGHEM M.
PROLOG, bases théoriques et développements actuels.
TSI, vol 2, n° 4, 1983.

COOPER W.S., MARON M.E.
Foundations of probabilistic utility theoric indexing.
Journal of the ACM, vol 25, n°1, janvier 1978.

COURTIN J.
Algorithmes pour le traitement interactif des langues naturelles.
Thèse d'Etat, Grenoble 1977.

COURTIN J., DUJARDIN D.
Paramètres linguistiques de la morphologie française dans le système P.I.A.F.
Document interne, Grenoble, décembre 1976.

COURTIN J., DUJARDIN D., KOWARSKI I., GENTHIAL D., STRUBE DE LIMA V.
Interactive multi-level systems for correction of ill-formed french texts.
Second Scandinavian Conference on Artificial Intelligence, Tempere, Finlande, juin 1989.

CROFT B.
An expert assistant for a document retrieval system.
RIAO 85, Grenoble, mars 1985.

CROFT B.
User specified domain knowledge for document retrieval.
ACM SIGIR, Pise, septembre 1986.

DEBILI F.
Traitements syntaxiques utilisant des matrices de précédence fréquentielles construites automatiquement par apprentissage.
Thèse de Docteur-Ingénieur, Paris VII, 1977.

DEBILI F.
Analyse syntaxico-sémantique fondée sur une acquisition automatique de relations lexicales-sémantiques.
Thèse d'Etat, Université de Paris XI, centre d'Orsay, 1982.

DEFUDE B.
Knowledge based system versus thesaurus: an architecture problem about expert system design.
ACM and BCS symposium, Research and development in information retrieval, Cambridge, juillet 1984.

DEFUDE B.
Etude et réalisation d'un système intelligent de recherche d'informations: Le prototype IOTA.
Thèse de l'INPG, Grenoble 1986.

DESCLES J.P.
Langages quasi-naturels articulés avec une base de connaissances: présentation et problèmes.
Colloque Traitement automatique des langues naturelles et systèmes documentaires, Clermont-Ferrand, mai 1982.

DIVAY M.
De l'écrit vers l'oral ou contribution à l'étude des traitements des textes écrits en vue de leur prononciation sur synthétiseur de parole.
Thèse d'Etat, Université de Rennes I, 1984.

FILLMORE C.J.
The case for case.
Universals in linguistic theory, Bach & Harms eds, Holt, Rinehart & Winston inc., 1968.

FLUHR C.
Algorithmes à apprentissage et traitement automatique des langues.
Thèse d'état, Université de Paris Sud, centre d'Orsay, 1977.

FLUHR C.
SPIRIT un système syntaxique et probabiliste d'indexation et de recherche d'informations textuelles.
IDT 81, Versailles, 1981.

FLUHR C., DEBILI F.
Interrogation en langue naturelle de données textuelles et factuelles.
RIAO 85, Grenoble, mars 1985.

GALIOTOU H.
Construction d'un analyseur morphologique du fançais.
Rapport de DEA, IMSS, Grenoble II, 1983.

GRANDJEAN E.
Algorithmes de construction et de mise à jour d'un dictionnaire.
Rapport de recherche n° 28, Grenoble 1976.

GRANDJEAN E., VEILLON G.
Utilisation d'une composante linguistique dans les logiciels de recherche d'informations: le logiciel prototype PIAFDOC.
Journées AFCET, Paris, avril 1980.

GREVISSE M.
Le bon usage.
Hatier, Paris dernière éd. 1980.

GROSS M.
Méthodes en syntaxe. Régime des constructions complétives.
Hermann, Paris, 1975.

GROSS M.
Lexicon Grammar - The representation of compound words.
COLING 86, Bonn, août 1986.

LE GUERN M.
Les descripteurs d'un système de documentation. Essai de définition.
Colloque Traitement automatique des langues naturelles et systèmes documentaires, Clermont-Ferrand, mai 1982.

HARRIS Z.
String analysis of sentence structure.
La Haye, Mouton, 1961.

HARRIS Z.
Structures mathématiques du langage.
Paris, Dunod, 1971.

HATON J.P.
Intelligence artificielle en compréhension automatique de la parole.
TSI, vol 4, n° 3, 1985.

HLAL Y.
Méthodes d'apprentissage pour l'analyse morpho-syntaxique (expérimentées dans le cas de l'arabe et du français).
Thése d'état, Orsay, 1977.

JAYEZ J.H.
Une approche de la compréhension par machine du langage naturel.
Thèse d'état, Paris VII,1979.

JAYEZ J.H.
Compréhension automatique du langage naturel. Le cas de l'interrogation simple en français.
Masson, Paris, 1986.

JIMENEZ GUARIN C.
Recherche par le contenu de textes structurés dans un environnement bureautique.
Rapport de DEA, INPG, 1985.

JIMENEZ GUARIN C.
Access by content of documents in an office information system.
ACM SIGIR, Grenoble, juin 1988.

JIMENEZ GUARIN C.

Opérations d'accès par le contenu à une base de documents textuels. Application à un environnement de bureau.
Thèse de l'INPG, juillet 1989.

JOLOBOFF V.

Unification d'arborescences. Evaluation sémantique d'énoncés en langue naturelle.
Thèse de Docteur-Ingénieur, Grenoble, 1978.

KALLAS G.

Résolution des solutions multiples en analyse morphologique automatique des langues naturelles. Utilisation des modèles de Markov.
Thèse de l'Université de Grenoble II, 1987.

KATZ J.J., FODOR J.A.

The structure of semantic theory.
in Fodor J.A., Katz J.J., "The structure of language", Prentice-Hall, 1964.

KAY M.

Morphological and syntactic analysis.
in A Zampolli, Linguistic structures processing, Amsterdam, North Holland, 1977.

KERKOUBA D.

Incidence du thésaurus dans les systèmes documentaires.
Rapport de DEA, Grenoble 1981.

KERKOUBA D.

Indexation et propriétés structurelles de documents dans un système de recherche d'informations.
Thèse de Docteur-Ingénieur, Grenoble 1984.

KERKOUBA D.

Automatic indexing and stuctural properties of texts.
RIAO 85, Grenoble, mars 1985.

KERKOUBA D., BRUANDET M.F.
Automatic indexing method using a partial knowledge model for an information retrieval system.
Document interne, janvier 1986.

LALLICH-BOIDIN G.
Analyse syntaxique automatique du français. Application à l'indexation automatique.
Thèse de l'Université de Grenoble II, 1986.

MAEEGAARD B., SPANG-HANSSEN E.
La segmentation automatique du français écrit.
Documents de Linguistique Quantitative n° 35, Dunod, 1978.

MELLISH C.S.
Computer interpretation of natural language descriptions.
Ellis Harwood series Artificial intelligence, 1985.

MEL'TCHUK I.A.
Linguistics, computational linguistics and Meaning-Text models.
Conférence ICCL, Pise, 1973.

MERLE A.
Un analyseur pré-syntaxique pour la levée des ambiguïtés dans des documents écrits en langue naturelle: application à l'indexation automatique.
Thèse de Docteur Ingénieur, Grenoble 1982.

MISTRAL
Manuel d'utilisation.
CII-HB, La documentation française, 1978.

NIE J.Y.
Compréhension de requêtes en langue naturelle. Pré-étude dans le contexte d'un système de recherche d'informations.
Rapport de DEA, Grenoble, 1981.

NIE J.Y.
Un modèle logique général pour les Systèmes de Recherche d'Informations. Application au prototype RIME.
Thèse de l'Université Joseph Fourier, Grenoble I, juillet 1990.

PALMER P.
Etude de l'organisation d'un dictionnaire pour l'analyse du français.
Rapport de DEA, Grenoble, 1981.

PALMER P., BERRUT C.
Etude d'un analyseur de surface de la langue naturelle pour un système de recherche documentaire.
13th CAIS Conference, Montréal, juin 1985.

PASSAT
Manuel de description.
SIEMENS, 1972.

PERENNOU G., DAUBEZE P., LAHENS F.
La vérification et la correction automatique de textes: le système VORTEX.
TSI, vol 5, n° 4, 1986.

PIERREL J.M.
Utilisation de contraintes linguistiques en compréhension automatique de la parole continue: le système MYRTILLE II.
TSI, vol 1, n° 5, 1982.

PITRAT J.
Textes, ordinateurs et compréhension.
Eyrolles, Paris, 1985.

QUILLAN M.R.
Semantic memory.
in Semantic Information Processing, Minsky M. ed., Cambridge, Mass, M.I.T. Press, 1968.

RADY M.
L'ambiguïté du langage naturel est-elle la source du non-déterminisme des procédures de traitement ?
Thèse d'Etat, Paris VI, 1983.

VAN RIJSBERGEN C.J.
Information retrieval.
Second edition, Buttersworth, London, 1979.

ROBINSON J.A.
A machine-oriented logic based on the resolution principle.
Journaf of the ACM 12, 1965.

ROUAULT J.
Linguistique automatique et informatique documentaire.
Colloque franco-anglais, DBMIST, Paris, décembre 1983.

SABAH G.
L'Intelligence Artificielle et le Language : Représentations des connaissances.
Editions Hermès, Paris, 1988.

SAGER N.
Syntactic Analysis of Natural Language.
in Advances in Computers, New-York, 1967.

SALKOFF M.
Une grammaire en chaî
ne du français. Analyse distributionnelle.
Dunod, Paris, 1973.

SALKOFF M.
Analyse syntaxique du français: Grammaire en chaîne.
Linguisticæ investigationes: supplementa, vol 2, John Benjamins B.V., Amsterdam, 1979.

SALKOFF M.
Analyse syntaxique.
15ème école de printemps d'informatique théorique, Informatique et Lingistique, Oléron, mai 1987.

SALTON G.
The SMART retrieval system - Experiment in automatic document processing.
Prentice-Hall, Englewood Cliffs, New Jersey, 1971.

SALTON G.
Recent tends in automatic information retrieval.
ACM SIGIR, Pise, septembre 1986.

SALTON G.
Historical note: the past thirty years in information retrieval.
Journal of the ASIS 38, 1987.

SALTON G., MC GILL M.J.
Introduction to modern information retrieval.
Mcgraw Hill book company, New-York, 1983.

SAMPSON G.
A stochastic Approach to Parsing.
COLING 86, Bonn, RFA, 1986.

SCHANK R.C.
Conceptual dependency, a theory of natural language understanding.
in Cognitive psychology, vol 3, n° 4, 1972.

SCHANK R.C., ABELSON R.
Scripts, plan and knowledge.
4th international joint conference on international intelligence. Tbilissi, août 1975.

SEDOGBO C.
Evaluation comparative de méthodes d'analyse syntaxique partielle.
Thèse de 3ème cycle, Paris, 1983.

SHEIL B.A.
Median split tree: a fast lookup technique for frequently occuring keys.
Communications of the ACM, novembre 1978.

SIMMONS R.F., SLOCUM J.
Generating english discourse from semantics nets.
Communications of the ACM, vol 15, 1972.

SMEATON A.F.
Incorporating syntactic information into a document retrieval stategy: an investigation.
ACM SIGIR, Pise, septembre 1986.

SMEATON A.F.
Experiments on incorporating syntactic processing of user queries into a document retrieval strategy.
ACM SIGIR, Genoble, juin 1988.

SMITH L.C.
Artificial intelligence applications in information systyems
Ann. Rev. Inform. Sci. Technol. 15, 1980.

SPARK JONES K.
Artificial intelligence: what can it offer to information retrieval ?
in Informatics 3, Aslib, London, 1978.

SPARK JONES K.
Computational Linguistics.
15ème école de printemps d'informatique théorique, Informatique et Lingistique, Oléron, mai 1987.

SPARK JONES K., WILKS Y.
Automatic natural language parsing.
Ellies Harwood series Artificial Intelligence, Chichester, 1983.

TESNIERES L.
Eléments de syntaxe structurale.
Klincksieck, Paris, 1959.

T.L.F.
Dictionnaire des fréquences.
CNRS T.L.F., Nancy, 1971.

TUTESCU M.
Le groupe nominal et la normalisation en français moderne.
Klincksieck, Paris, 1972.

VAUQUOIS B.
La traduction automatique à Grenoble.
Document de Linguistique Quantitative n° 24, Dunod, 1975.

VAUQUOIS B.
Bernard Vauquois et la TAO, vingt-cinq ans de Traduction Automatique, ANALECTES.
Ass. Champollion & GETA, Grenoble, 1988

VERGNE J.
Analyse morpho-syntaxique automatique sans dictionnaire.
Thèse de l'Université Paris 6, Paris, 1989.

WEIZENBAUM J.
ELIZA, a computer program for the study of natural language communication between man and machine.
Communication of the ACM, janvier 1966.

WINOGRAD T.
Understanding Natural Language.
in Cognitive Psychology, janvier 1972.

WOODS W.A.
Transition network grammars for natural language analysis.
Communication of the ACM, octobre 1970.

WOODS W.A.
Cascaded ATN Grammars.
American Journal of Computational Linguistics 6, 1980.

YAN J.
Interprétation sémantique des comptes rendus médicaux.
Rapport de DEA, INPG/USTMG, Grenoble 1986.

ZAJAC R.
Etude des possibilités d'interaction Homme-Machine dans un processus de traduction automatique.
Thèse de l'INPG, Grenoble, 1986.

ZHOLKOVSKIJ A.K., MEL'TCHUK I.A.
Sur la synthèse sémantique.
T.A. informations n° 2, 1970 (traduction du russe 1967).

ANNEXES

Annexe 1

Liste des catégories grammaticales

SUBC	substantif commun
SUBP	substantif propre
ADJQ	adjectif qualificatif
ADJC	adjectif numéral cardinal
ADJO	adjectif numéral ordinal
ADJP	adjectif possessif
ADJR	adjectif relatif
ADJD	adjectif démonstratif
ADJI	adjectif indéfini
ADJE	adjectif interrogatif / exclamatif
ARTD	article défini
ARTI	article indéfini
ARTC	article contracté
PRPS	pronom personnel sujet
PRPV	pronom personnel préverbal
PRPC	pronom personnel complément
PADV	pronom personnel adverbial
PRPO	pronom possessif
PRDM	pronom démonstratif
PRRL	pronom relatif
PRIN	pronom indéfini
PREP	préposition
CONJ	conjonction
ADVB	adverbe
XECJ	auxiliaire être conjugué
XEIF	auxiliaire être à l'infinitif
XEPA	auxiliaire être au participe passé
XEPR	auxiliaire être au participe présent
XACJ	auxiliaire avoir conjugué
XAIF	auxiliaire avoir à l'infinitif
XAPA	auxiliaire avoir au participe passé
XAPR	auxiliaire avoir au participe présent
VBCJ	verbe conjugué
VBIF	verbe à l'infinitif
VBPA	verbe au participe passé
VBPR	verbe au participe présent

ADVN	adverbe de négation "ne"
ADVP	adverbe de négation "pas"
LOCP	locution prépositive
LOCC	locution conjonctive
INTJ	interjection
NOMB	nombre
PNTF	ponctuation forte
PARE	parenthèses
OPER	opérateurs
VIRG	virgule
DPNT	deux-points
ABRV	abréviation et sigle

Annexe 2

Associations catégorie-variables grammaticales

SUBC	genre	nombre	
SUBP	genre	nombre	
ADJQ	genre	nombre	
ADJC	genre	nombre	
ADJO	genre	nombre	
ADJP	genre	nombre	
ADJR	genre	nombre	
ADJD	genre	nombre	
ADJI	genre	nombre	
ADJE	genre	nombre	
ARTD	genre	nombre	
ARTI	genre	nombre	
ARTC	genre	nombre	
PRPS	personne	nombre	
PRPV	genre	nombre	
PRPC	genre	nombre	
PRPO	genre	nombre	
PRDM	genre	nombre	
PRRL	genre	nombre	
PRIN	genre	nombre	
XECJ	mode	temps	personne nombre
XEPA	genre	nombre	
XACJ	mode	temps	personne nombre
XAPA	genre	nombre	
VBCJ	mode	temps	personne nombre
VBPA	genre	nombre	

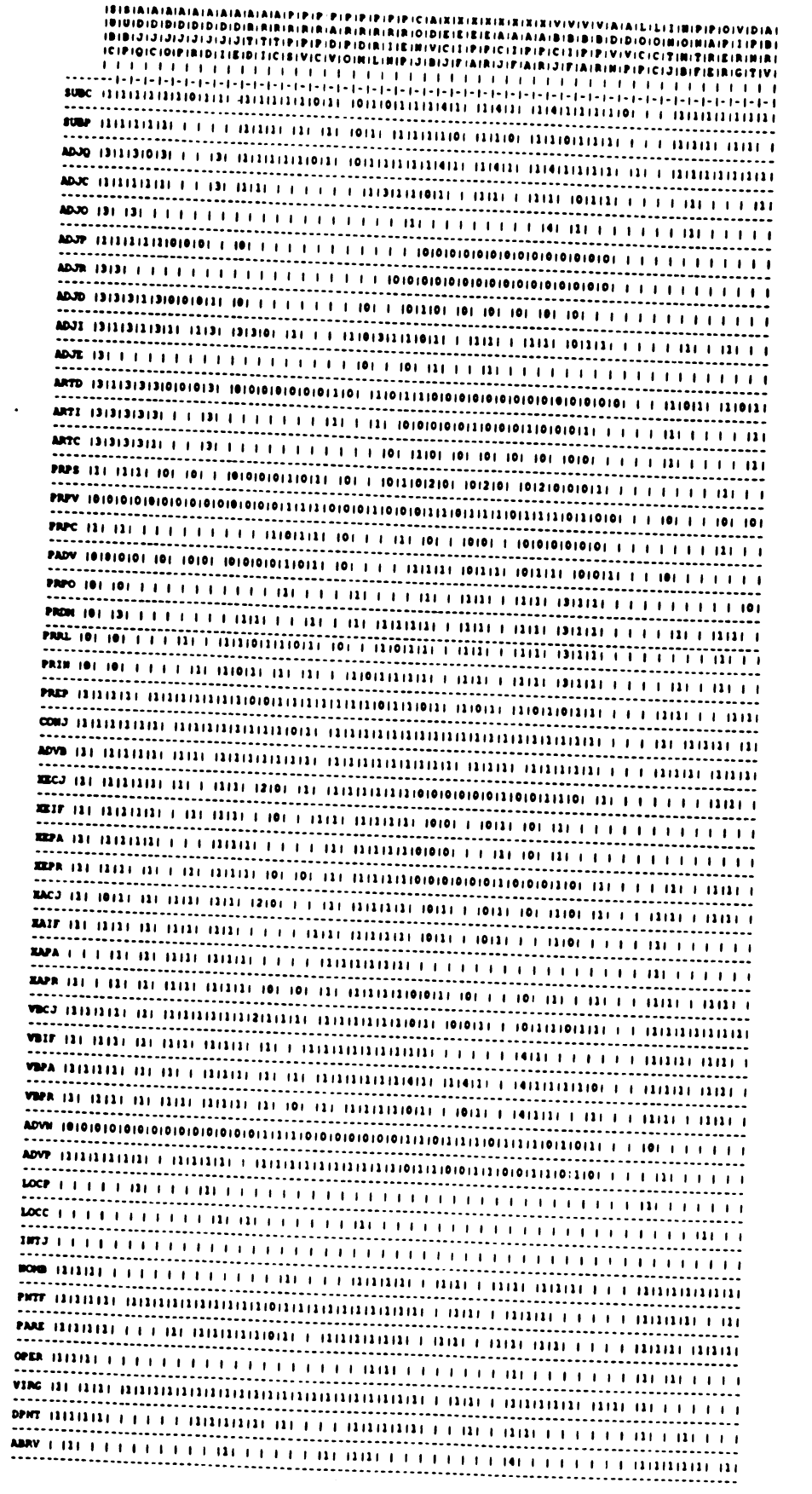
Annexe 3

Liste des symboles non terminaux de la grammaire des GC

GC	groupe conceptuel
CPV	complément de verbe
CPN	complément de nom
CP	complément
NQ	nom qualifié
GADJ	groupe adjectival
ARG	argument
NOM	nom (simple ou composé)
ADJ	adjectif
SUB	substantif
INF	infinitif
PREP	préposition
EP	épithète
AT	attribut
D	"de"
ND	non "de"
P	préposition autre que "de"

Annexe 4

Matrice de précédence



Annexe 5

catégorisation en fonction des terminaisons

a	-->	SUBC	{ masculin / féminin , singulier }
		ADJQ	{ masculin / féminin , singulier }
		VBCJ	{ 3ème , singulier }
é	-->	SUBC	{ masculin / féminin , singulier }
		ADJQ	{ masculin , singulier }
		VBPA	{ masculin , singulier }
ée , ie	-->	SUBC	{ masculin / féminin , singulier }
		ADJQ	{ masculin / féminin , singulier }
		VBPA	{ féminin , singulier }
		VBCJ	{ 1ère / 3ème , singulier }
≠(é,i)e	-->	SUBC	{ masculin / féminin , singulier }
		ADJQ	{ masculin / féminin , singulier }
		VBCJ	{ 1ère / 3ème , singulier }
ai	-->	SUBC	{ masculin , singulier }
		ADJQ	{ masculin , singulier }
		VBCJ	{ 1ère , singulier }
≠ai	-->	SUBC	{ masculin / féminin , singulier }
		ADJQ	{ masculin / féminin , singulier }
		VBCJ	{ 1ère , singulier }
		VBPA	{ masculin , singulier }
er	-->	SUBC	{ masculin / féminin , singulier }
		ADJQ	{ masculin / féminin , singulier }
		VBIF	
ir	-->	SUBC	{ masculin , singulier }
		ADJQ	{ masculin , singulier }
		VBIF	

≠(e,i)r	-->	SUBC	{ masculin / féminin , singulier / pluriel }
		ADJQ	{ masculin / féminin , singulier / pluriel }
ez	-->	SUBC	{ masculin , singulier / pluriel }
		VBCJ	{ 2ème , pluriel }
≠ez	-->	SUBC	{ masculin / féminin , singulier / pluriel }
		ADJQ	{ masculin / féminin , singulier / pluriel }
és	-->	SUBC	{ masculin / féminin , pluriel }
		ADJQ	{ masculin , pluriel }
		VBPA	{ masculin , pluriel }
is	-->	SUBC	{ masculin , singulier / pluriel }
		ADJQ	{ masculin , singulier / pluriel }
		VBCJ	{ 1ère / 2ème , singulier }
		VBPA	{ masculin , pluriel }
as	-->	SUBC	{ masculin / féminin , singulier / pluriel }
		ADJQ	{ masculin / féminin , singulier / pluriel }
		VBCJ	{ 2ème , singulier }
ées , ies	-->	SUBC	{ masculin / féminin , pluriel }
		ADJQ	{ masculin / féminin , pluriel }
		VBPA	{ féminin , pluriel }
		VBCJ	{ 1ère , singulier }
			{ 2ème , singulier / pluriel }
≠(é,i)es	-->	SUBC	{ masculin / féminin , pluriel }
		ADJQ	{ masculin / féminin , pluriel }
		VBCJ	{ 1ère , singulier }
			{ 2ème , singulier / pluriel }
ons	-->	SUBC	{ masculin / féminin , pluriel }
		ADJQ	{ masculin , pluriel }
		VBCJ	{ 1ère , pluriel }
≠ons	-->	SUBC	{ masculin / féminin , singulier / pluriel }
		ADJQ	{ masculin / féminin , singulier / pluriel }

≠(e,n)s	-->	SUBC	{ masculin / féminin , singulier / pluriel }
		ADJQ	{ masculin / féminin , singulier / pluriel }
ant	-->	SUBC	{ masculin , singulier }
		ADJQ	{ masculin , singulier }
		VBPR	
ont	-->	SUBC	{ masculin , singulier }
		ADJQ	{ masculin , singulier }
		VBCJ	{ 3ème , singulier / pluriel }
ment	-->	SUBC	{ masculin , singulier }
		ADJQ	{ masculin , singulier }
		VBCJ	{ 3ème , pluriel }
		ADVB	
≠ment	-->	SUBC	{ masculin / féminin , singulier }
		ADJQ	{ masculin , singulier }
		VBCJ	{ 3ème , pluriel }
≠(a,o,e)nt	-->	SUBC	{ masculin / féminin , singulier / pluriel }
		ADJQ	{ masculin / féminin , singulier / pluriel }
it , ât	-->	SUBC	{ masculin , singulier }
		ADJQ	{ masculin , singulier }
		VBCJ	{ 3ème , singulier / pluriel }
SINON	-->	SUBC	{ masculin / féminin , singulier / pluriel }
		ADJQ	{ masculin / féminin , singulier / pluriel }

Annexe 6

extraits des différents textes analysés

Cette annexe comprend des extraits des textes ayant servi à l'expérimentation présentée au chapitre IV. Nous donnons pour chacun de ces extraits les Groupes Conceptuels reconnus sans et avec regroupement, noté SAV, des catégories SUBC, ADJQ et VBPA. Chacune des formes participant à un GC est suivie de son rang dans le texte, de sa catégorie grammaticale, et de son numéro d'unité lexicale (un zéro signifie que la forme est inconnue). Ces extraits proviennent des textes suivants :

• crmed est un compte rendu médical (compte rendu radiologique rédigé par des médecins) extrait du corpus d'expérimentation de RIME. Ce compte rendu médical comporte près de 50% de formes inconnues.

• intro est l'introduction de ce mémoire. Pour ce texte le taux de reconnaissance du vocabulaire est de 78,2%.

• ch4 et ch13 sont deux chapitres du corpus des NEF (utilisé comme corpus d'expérimentation de IOTA). Pour ces 2 textes les taux de reconnaissance du vocabulaire sont respectivement de 81% et de 79,6%.

crmed

Les coupes ont {t{ pratiqu{es apr}s radiographie num{rique.
Images complexes associant
. opacit{ de la grande cavit{ pleurale droite de nature hydrique.
. opacit{ pulmonaire en projection du lobe sup{rieur droit et d'aspect a lv{olaire avec signe de n{crose.
. opacit{s pulmonaires contro-lat{rales, d'aspect nodulaire ou @ contour plus flou, hautement suspect dans le contexte actuel.
. opacit{ @ disposition p{ri-bronchique de 25 mm de diam}tre, en projection de la bronche lobaire sup{rieure droite et du tronc bronchique interm{diaire correspondant tr}s probablement @ la masse tumorale.
. extension ganglionnaire m{diastinale, au niveau du groupe de la bifurcation de la cha^ine para-trach{ale droite et de la cha^ine m{diastinale ant{rieure droite; apr}s injection intra-veineuse de produit de contraste certains ganglions sont le si}ge d'une n{crose massive.

Les constatations sont les suivantes:
. confirmation d'une hypertrophie de densit{ tissulaire, du lobe gauche du corps thyro-ide avec extension m{diane dans l'orifice cervico-thoracique.
. importante masse de si}ge p{diculaire gauche, avec extension pulmonaire et surtout m{diastinale sous l'aspect d'ad{nopathie, sup{rieure @ 20 mm de diam}tre, int{ressant notamment la cha^ine m{diastinale ant{rieure gauche.
. opacit{ ganglionnaire sup{rieure @ 10 mm au niveau de la partie haute de la cha^ine para-trach{ale droite.

extraction des Groupes Conceptuels sans regroupement

```
1

projection   25   SUBC  0

2

n(crose   37   SUBC  0

3

disposition   57   SUBC  3191
p(ri   58   VBPA  1669

4

diam)tre   64   SUBC  0

5

projection   66   SUBC  0

6

bifurcation   94   SUBC  0

7

produit   113   SUBC  2241
de   114   PREP  689
contraste   115   SUBC  0

8

si)ge   120   SUBC  0

9

suivantes   129   SUBC  2357

10

hypertrophie   133   SUBC  0

11

aspect   164   SUBC  0
d'   165   PREP  689
ad(nopathie   166   SUBC  0

12

diam)tre   172   SUBC  0

13

niveau   187   SUBC  0
de   188   PREP  689
la   189   ARTD  612
partie   190   SUBC  2185
```

extraction des Groupes Conceptuels avec regroupement

```
1

images   9   SAV   0
complexes   10   SAV   0

2

opacit(   12   SAV   0

3

cavit(   16   SAV   0

4

nature   20   SAV   3218

5

opacit(   22   SAV   0

6

projection   25   SUBC   0
du   26   ARTC   1972
lobe   27   SAV   0
sup(rieur   28   SAV   2359

7

aspect   32   SAV   0

8

n(crose   37   SUBC   0

9

opacit(s   38   SAV   0
pulmonaires   39   SAV   0
contro   40   SAV   0
lat(rales   41   SAV   0

10

aspect   43   SAV   0

11

contour   47   SAV   0

12

flou   49   SAV   0

13

suspect   51   SAV   0
dans le   52   ARTC   9
contexte   53   SAV   0
actuel   54   SAV   0

14

opacit(   55   SAV   0
@   56   PREP   685
disposition   57   SUBC   3191
p(ri   58   VBPA   1669

15

diam)tre   64   SUBC   0

16

projection   66   SUBC   0
de   67   PREP   689
la   68   ARTD   612
bronche   69   SAV   0

17

sup(rieure   71   SAV   2359

18

tronc   75   SAV   0

19

masse   83   SAV   0

20

extension   85   SAV   0

21

bifurcation   94   SUBC   0
de   95   PREP   689
la   96   ARTD   612
cha^ine   97   SAV   0

22

cha^ine   104   SAV   0

23

injection   109   SAV   0

24

produit   113   SUBC   2241
de   114   PREP   689
contraste   115   SUBC   0

25

ganglions   117   SAV   0

26

si)ge   120   SUBC   0
d'   121   PREP   689
une   122   ARTI   613
n(crose   123   SAV   0

27

constatations   126   SAV   0

28

suivantes   129   SUBC   2357

29

confirmation   130   SAV   0
d'   131   PREP   689
une   132   ARTI   613
hypertrophie   133   SUBC   0
de   134   PREP   689
densit(   135   SAV   0

30

lobe   138   SAV   0

31

corps   141   SAV   0

32

extension   144   SAV   0

33

orifice   147   SAV   0
cervico   148   SAV   0

34

importante   150   SAV   0

35

si)ge   153   SAV   0

36

extension   157   SAV   0
```

```
37

aspect   164   SUBC  0
d'   165   PREP  689
ad(nopathie   166   SUBC  0

38

sup(rieure   167   SAV   2359

39

diam)tre   172   SUBC  0

40

opacit(   180   SAV   0

41

sup(rieure   182   SAV   2359

42

niveau   187   SUBC  0
de   188   PREP  689
la   189   ARTD  612
partie   190   SUBC  2185

43

cha^ine   194   SAV   0
```

intro

Ce travail a {t{ r{alis{ dans le cadre du projet IOTA (d{veloppement d'un prototype de syst)me intelligent de recherche d'informations) au sein de l'{quipe SIRI (Syst)mes Intelligents de Recherche d'Informations) du laboratoire de G{nie Informatique de Grenoble,
et a pour objet la r{alisation d'un module d'analyse linguistique destin{ @ permettre une indexation automatique de textes en langue naturelle.

La probl)matique de la recherche d'informations consiste @ partir de l'expression d'une requ^ete d'un utilisateur,
qui d{finit le th)me de la recherche, @ retrouver un ensemble de documents (ou de r{f{rences @ des documents) dont le contenu s{mantique correspond le mieux au th)me recherch{, parmi l'ensemble des documents de la base documentaire consid{r{e.
Cette recherche s'effectue par l'exploitation d'une fonction de correspondance qui {tablit une association entre un th)me de recherche (requ^ete) et un contenu s{mantique (documents).
Cette fonction de correspondance est fond{e sur la d{finition de mod)les s{mantiques pour les th)mes et les documents, et sur la d{finition de crit)res de correspondance entre ces {l{ments [CHIA 88].
Nous pouvons sch{matiser la probl)matique des syst)mes de recherche d'informations (SRI), de la mani)re suivante:

La repr{sentation d'un document dans le mod)le s{mantique correspondant est r{alis{e par la fonction d'indexation du syst)me de recherche d'informations, la compr{hension de la requ^ete et l'{valuation de la fonction de correspondance sont r{alis{es par la fonction d'interrogation du SRI.
Nous pr{sentons les diff{rents composants des fonctions d'indexation et d'interrogation d'un syst)me de recherche d'informations dans le premier chapitre.
L'objectif d'un module d'analyse linguistique pour la r{alisation de la fonction d'indexation d'un SRI, est de permettre la repr{sentation du contenu s{mantique des documents,
par la reconnaissance automatique, @ partir de leur contenu textuel, de concepts structur{s (mots ou groupes de mots) susceptibles de constituer des {l{ments du mod)le s{mantique.
Nous nous int{ressons ici @ un mod)le s{mantique simple dans son principe: le contenu de chaque document est repr{sent{ par un ensemble de termes (appel{s termes d'indexation), qui correspondent @ autant de concepts significatifs du contenu s{mantique du document.
Le degr{ d'{laboration du mod)le s{mantique des documents qui conditionne l'efficacit{ de la fonction de correspondance et donc l'efficacit{ du SRI,
est donc directement d{pendant du niveau de repr{sentation des termes d'indexation.
Ce niveau pouvant ^etre plus ou moins proche de celui du langage naturel: mots isol{s, syntagmes phrases ...

Nous nommons Groupes Conceptuels (G.C.) la repr{sentation normalis{e, dans un formalisme cible, de ces concepts structur{s qui correspondent dans IOTA aux groupes nominaux.
La pr{sentation des groupes conceptuels et des diff{rents choix ayant conduit @ leur d{finition est effectu{e dans le troisi)me chapitre.

extraction des Groupes Conceptuels sans regroupement

```
1

Ce   4   SUBP  0

2

cadre   10   SUBC  0

3

prototype   17   SUBC  2725

4

informations   24   SUBC  3245

5

sein   26   SUBC  0

6

Syst)mes   31   SUBP  0
Intelligents   32   SUBP  0
de   33   PREP  689
Recherche   34   SUBP  0
d'   35   PREP  689
Informations   36   SUBP  0

7

laboratoire   38   SUBC  0
de   39   PREP  689
G(nie   40   SUBP  0
Informatique   41   SUBP  0
de   42   PREP  689
Grenoble   43   SUBP  0

8

objet   47   SUBC  3212

9

r(alisation   49   SUBC  2724
d'   50   PREP  689
un   51   ARTI  613
module   52   SUBC  0

10

permettre   58   VBIF  211

11

textes   63   SUBC  2791

12

probl)matique   68   SUBC  0

13

informations   73   SUBC  3245

14

partir   76   VBIF  51
de   77   PREP  689
l'   78   ARTD  612
expression   79   SUBC  0
d'   80   PREP  689
une   81   ARTI  613
requ^ete   82   SUBC  0

15

th)me   89   SUBC  0

16

retrouver   94   VBIF  1547

17

ensemble   96   SUBC  625
de   97   PREP  689
documents   98   SUBC  3223

18

r(f(rences   101   SUBC  2720
@   102   PREP  685
des   103   ARTI  613
documents   104   SUBC  3223

19

contenu   107   SUBC  1932

20

documents   119   SUBC  3223

21

exploitation   131   SUBC  3241
d'   132   PREP  689
une   133   ARTI  613
fonction   134   SUBC  0

22

association   140   SUBC  0

23

contenu   149   SUBC  1932

24

documents   151   SUBC  3223

25

fonction   153   SUBC  0

26

d(finition   160   SUBC  0
de   161   PREP  689
mod)les   162   SUBC  2141

27

documents   169   SUBC  3223

28

d(finition   173   SUBC  0
de   174   PREP  689
crit)res   175   SUBC  0
de   176   PREP  689
correspondance   177   SUBC  0

29

informations   194   SUBC  3245

30

La   200   SUBP  0

31

document   204   SUBC  3223
dans le   205   ARTC  9
mod)le   206   SUBC  2141

32

fonction   213   SUBC  0
d'   214   PREP  689
indexation   215   SUBC  0

33

informations   221   SUBC  3245

34

compr(hension   223   SUBC  0

35

(valuation   229   SUBC  0
de   230   PREP  689
la   231   ARTD  612
fonction   232   SUBC  0
```

```
36

fonction   239   SUBC  0
d'   240   PREP  689
interrogation   241   SUBC  0

37

fonctions   250   SUBC  0

38

interrogation   255   SUBC  0

39

informations   262   SUBC  3245

40

objectif   267   SUBC  0
d'   268   PREP  689
un   269   ARTI  613
module   270   SUBC  0

41

r{alisation   276   SUBC  2724
de   277   PREP  689
la   278   ARTD  612
fonction   279   SUBC  0
d'   280   PREP  689
indexation   281   SUBC  0

42

permettre   287   VBIF  211

43

contenu   291   SUBC  1932

44

documents   294   SUBC  3223

45

partir   300   VBIF  51

46

{l{ments   317   SUBC  0
du   318   ARTC  1972
mod}le   319   SUBC  2141

47

mod}le   327   SUBC  2141

48

contenu   334   SUBC  1932

49

document   337   SUBC  3223

50

ensemble   342   SUBC  625

51

contenu   357   SUBC  1932

52

document   360   SUBC  3223

53

degr{   362   SUBC  0
d'   363   PREP  689
{laboration   364   SUBC  0
du   365   ARTC  1972
mod}le   366   SUBC  2141

54

documents   369   SUBC  3223

55

efficacit{   373   SUBC  0
de   374   PREP  689
la   375   ARTD  612
fonction   376   SUBC  0

56

efficacit{   382   SUBC  0

57

niveau   390   SUBC  0
de   391   PREP  689
repr{sentation   392   SUBC  0
des   393   ARTC  1972
termes   394   SUBC  0
d'   395   PREP  689
indexation   396   SUBC  0

58

langage   406   SUBC  2770

59

Groupes   414   SUBP  0
Conceptuels   415   SUBP  0

60

G   416   SUBP  0

61

pr{sentation   437   SUBC  2766
```

extraction des Groupes Conceptuels avec regroupement

```
1

Ce   4   SUBP  0
travail   5   SAV   0

2

cadre   10   SUBC  0
du   11   ARTC  1972
projet   12   SAV   0

3

prototype   17   SUBC  2725
de   18   PREP  689
syst)me   19   SAV   0

4

informations   24   SUBC  3245

5

sein   26   SUBC  0
de   27   PREP  689
l'   28   ARTD  612
(quipe   29   SAV   0

6

Syst)mes   31   SUBP  0
Intelligents   32   SUBP  0
de   33   PREP  689
Recherche   34   SUBP  0
d'   35   PREP  689
Informations   36   SUBP  0

7

laboratoire   38   SUBC  0
de   39   PREP  689
G(nie   40   SUBP  0
Informatique   41   SUBP  0
de   42   PREP  689
Grenoble   43   SUBP  0

8

objet   47   SUBC  3212

9

r(alisation   49   SUBC  2724
d'   50   PREP  689
un   51   ARTI  613
module   52   SUBC  0
d'   53   PREP  689
analyse   54   SAV   0

10

destin(   56   SAV   0
@   57   PREP  685
permettre   58   VBIF  211

11

indexation   60   SAV   0

12

textes   63   SUBC  2791

13

probl)matique   68   SUBC  0

14

informations   73   SAV   3245

15

partir   76   VBIF  51
de   77   PREP  689
l'   78   ARTD  612
expression   79   SUBC  0
d'   80   PREP  689
une   81   ARTI  613
requ^ete   82   SUBC  0
d'   83   PREP  689
un   84   ARTI  613
utilisateur   85   SAV   0

16

th)me   89   SUBC  0

17

retrouver   94   VBIF  1547

18

ensemble   96   SUBC  625
de   97   PREP  689
documents   98   SUBC  3223

19

r(f(rences   101   SUBC  2720
@   102   PREP  685
des   103   ARTI  613
documents   104   SUBC  3223

20

contenu   107   SUBC  1932
s(mantique   108   SAV   0

21

th)me   113   SAV   0
recherch(   114   VBPA  1502

22

documents   119   SUBC  3223
de   120   PREP  689
la   121   ARTD  612
base   122   SAV   0

23

exploitation   131   SUBC  3241
d'   132   PREP  689
une   133   ARTI  613
fonction   134   SUBC  0
de   135   PREP  689
correspondance   136   SAV   0

24

association   140   SUBC  0

25

contenu   149   SUBC  1932

26

documents   151   SUBC  3223

27

fonction   153   SUBC  0
de   154   PREP  689
correspondance   155   SAV   0

28

d(finition   160   SUBC  0
de   161   PREP  689
mod)les   162   SUBC  2141
s(mantiques   163   SAV   0
pour   164   PREP  697
les   165   ARTD  612
th)mes   166   SAV   0
```

29

documents 169 SUBC 3223

30

d(finition 173 SUBC 0
de 174 PREP 689
crit)res 175 SUBC 0
de 176 PREP 689
correspondance 177 SUBC 0
entre 178 PREP 692
ces 179 ADJD 607
(l(ments 180 SAV 0

31

informations 194 SUBC 3245

32

mani)re 198 SAV 0
suivante 199 SAV 2357

33

La 200 SUBP 0
repr(sentation 201 SAV 0
d' 202 PREP 689
un 203 ARTI 613
document 204 SUBC 3223
dans le 205 ARTC 9
mod)le 206 SUBC 2141

34

fonction 213 SUBC 0
d' 214 PREP 689
indexation 215 SUBC 0

35

informations 221 SUBC 3245

36

compr(hension 223 SUBC 0
de 224 PREP 689
la 225 ARTD 612
requ^ete 226 SAV 0

37

(valuation 229 SUBC 0
de 230 PREP 689
la 231 ARTD 612
fonction 232 SUBC 0
de 233 PREP 689
correspondance 234 SAV 0

38

fonction 239 SUBC 0
d' 240 PREP 689
interrogation 241 SUBC 0

39

composants 248 SAV 0
des 249 ARTC 1972
fonctions 250 SUBC 0
d' 251 PREP 689
indexation 252 SAV 0

40

interrogation 255 SUBC 0

41

informations 262 SUBC 3245

42

objectif 267 SUBC 0
d' 268 PREP 689
un 269 ARTI 613
module 270 SUBC 0
d' 271 PREP 689
analyse 272 SAV 0

43

r(alisation 276 SUBC 2724
de 277 PREP 689
la 278 ARTD 612
fonction 279 SUBC 0
d' 280 PREP 689
indexation 281 SUBC 0

44

permettre 287 VBIF 211

45

repr(sentation 289 SAV 0
du 290 ARTC 1972
contenu 291 SUBC 1932

46

documents 294 SUBC 3223

47

reconnaissance 297 SAV 0

48

partir 300 VBIF 51

49

contenu 303 SAV 1932
textuel 304 SAV 0

50

concepts 306 SAV 0
structur(s 307 VBPA 2682

51

mots 308 SAV 0

52

mots 312 SAV 0

53

(l(ments 317 SUBC 0
du 318 ARTC 1972
mod)le 319 SUBC 2141

54

mod)le 327 SUBC 2141

55

contenu 334 SUBC 1932

56

document 337 SUBC 3223

57

ensemble 342 SUBC 625
de 343 PREP 689
termes 344 SAV 0

58

appel(s 345 VBPA 1798
termes 346 SAV 0
d' 347 PREP 689
indexation 348 SAV 0

59

concepts 354 SAV 0
significatifs 355 SAV 0
du 356 ARTC 1972
contenu 357 SUBC 1932

60

document 360 SUBC 3223

```
61

degr(   362   SUBC  0
d'   363   PREP  689
(laboration   364   SUBC  0
du   365   ARTC  1972
mod)le   366   SUBC  2141

62

documents   369   SUBC  3223

63

efficacit(   373   SUBC  0
de   374   PREP  689
la   375   ARTD  612
fonction   376   SUBC  0
de   377   PREP  689
correspondance   378   SAV   0

64

efficacit(   382   SUBC  0

65

niveau   390   SUBC  0
de   391   PREP  689
repr(sentation   392   SUBC  0
des   393   ARTC  1972
termes   394   SUBC  0
d'   395   PREP  689
indexation   396   SUBC  0

66

niveau   398   SAV   0

67

langage   406   SUBC  2770
naturel   407   SAV   0

68

mots   408   SAV   0
isol(s   409   VBPA  1410

69

phrases   411   SAV   0

70

Groupes   414   SUBP  0
Conceptuels   415   SUBP  0

71

G   416   SUBP  0

72

repr(sentation   419   SAV   0

73

formalisme   423   SAV   0

74

concepts   427   SAV   0
structur(s   428   VBPA  2682

75

nominaux   435   SAV   0

76

pr(sentation   437   SUBC  2766

77

conceptuels   440   SAV   0

78

choix   444   SAV   0

79

conduit   446   SAV   1923

80

d(finition   449   SAV   0
```

ch4

Services sp{ciaux.

G{n{ralit{s.

La num{rotation des services sp{ciaux pourra ^etre modifi{e (cf.nouveau plan de num{rotage).
Chacun des services pourra ^etre desservi par un faisceau direct ou par l'interm{diaire d'un centre de transit.
La commande du passage du faisceau direct au faisceau empruntant un centre de transit,et vice versa,pourra se faire :
soit par commande manuelle,

soit par commande par une horloge incorpor{e @ l'autocommutateur.

services manuels obtenus par le pr{fixe 10.

Les abonn{s d{sirant une communication manuelle au d{part d'un autocommutateur l'obtiennent en composant le 10 (ou le 15 en province).

Les communications ainsi obtenues sont tax{es par ticket.

Les autocommutateurs n'auront pas @ se pr{occuper de la taxation des communications demand{es par voie manuelle.
Les signaux de ligne utilis{s devront ^etre ceux du code de signalisation pour l'acc}s @ l'interurbain manuel (10).
services obtenus par le pr{fixe 12.

Les appels @ destination du service des renseignements t{l{phoniques,ou de tout autre service desservi par le pr{fixe 12,devront pouvoir ^etre achenim{s :

par un faisceau direct,

par l'interm{diaire d'un centre de transit,

si le service est local,par des lignes d'abonn{s,(ventuellement constitu{es en groupement.
Ces acheminements devront pouvoir coexister dans un m^eme autocommutateur,dans le cas o| le 12 est desservi par des faisceaux distincts suivant les heures.

Dans le cas d'un acheminement par faisceau direct :

on ne devra pas utiliser de signaux d'enregistreurs,

les signaux de ligne utilis{s devront ^etre ceux du code de signalisation pour l'acc}s au service des renseignements (12).
Lorsque l'acheminement est effectu{ par l'interm{diaire d'un centre de transit,la signalisation utilis{e doit ^etre le code MF SOCOTEL (signaux de ligne et d'enregistreurs).

extraction des Groupes Conceptuels sans regroupement

```
1

services   4   SUBC  2702
sp(ciaux   5   ADJQ  2701

2

g(n(ralit(s   6   SUBC  0

3

num(rotation   8   SUBC  0
des   9   ARTC  1972
services   10   SUBC  2702
sp(ciaux   11   ADJQ  2701

4

services   22   SUBC  2702

5

interm(diaire   33   SUBC  0
d'   34   PREP  689
un   35   ARTI  613
centre   36   SUBC  0
de   37   PREP  689
transit   38   SUBC  0

6

passage   42   SUBC  0

7

centre   50   SUBC  0
de   51   PREP  689
transit   52   SUBC  0

8

faire   57   VBIF  254

9

autocommutateur   71   SUBC  0

10

services   72   SUBC  2702

11

d(part   86   SUBC  2002
d'   87   PREP  689
un   88   ARTI  613
autocommutateur   89   SUBC  0

12

ticket   108   SUBC  0

13

pr(occuper   116   VBIF  1476
de   117   PREP  689
la   118   ARTD  612
taxation   119   SUBC  0

14

voie   124   SUBC  2431

15

signaux   127   SUBC  0
de   128   PREP  689
ligne   129   SUBC  2798
utilis(s   130   VBPA  2683

16

code   135   SUBC  3260
de   136   PREP  689
signalisation   137   SUBC  0
pour   138   PREP  697
l'   139   ARTD  612
acc)s   140   SUBC  0

17

services   146   SUBC  2702
obtenus   147   VBPA  21

18

appels   153   SUBC  0
@   154   PREP  685
destination   155   SUBC  0
du   156   ARTC  1972
service   157   SUBC  2702
des   158   ARTC  1972
renseignements   159   SUBC  3236

19

service   165   SUBC  2702
desservi   166   VBPA  91

20

pouvoir   172   VBIF  116

21

interm(diaire   181   SUBC  0
d'   182   PREP  689
un   183   ARTI  613
centre   184   SUBC  0
de   185   PREP  689
transit   186   SUBC  0

22

service   189   SUBC  2702

23

lignes   194   SUBC  2798
d'   195   PREP  689
abonn(s   196   SUBC  0

24

pouvoir   204   VBIF  116
coexister   205   VBIF  0

25

cas   211   SUBC  3255

26

cas   225   SUBC  3255
d'   226   PREP  689
un   227   ARTI  613
acheminement   228   SUBC  0

27

utiliser   236   VBIF  2683
de   237   PREP  689
signaux   238   SUBC  0
d'   239   PREP  689
enregistreurs   240   SUBC  0

28

signaux   242   SUBC  0
de   243   PREP  689
ligne   244   SUBC  2798
utilis(s   245   VBPA  2683

29

code   250   SUBC  3260
de   251   PREP  689
signalisation   252   SUBC  0
pour   253   PREP  697
l'   254   ARTD  612
acc)s   255   SUBC  0
au   256   ARTC  1860
service   257   SUBC  2702
des   258   ARTC  1972
renseignements   259   SUBC  3236
```

```
30

interm(diaire   268   SUBC  0
d'   269   PREP  689
un   270   ARTI  613
centre   271   SUBC  0
de   272   PREP  689
transit   273   SUBC  0

31

code   280   SUBC  3260

32

ligne   285   SUBC  2798
```

extraction des Groupes Conceptuels avec regroupement

```
1

services   4   SUBC  2702
sp(ciaux   5   ADJQ  2701

2

g(n(ralit(s   6   SUBC  0

3

num(rotation   8   SUBC  0
des   9   ARTC  1972
services   10   SUBC  2702
sp(ciaux   11   ADJQ  2701

4

nouveau   16   ADJQ  424
plan   17   SAV   0
de   18   PREP  689
num(rotage   19   SAV   0

5

services   22   SUBC  2702

6

faisceau   28   SAV   0
direct   29   SAV   0

7

interm(diaire   33   SUBC  0
d'   34   PREP  689
un   35   ARTI  613
centre   36   SUBC  0
de   37   PREP  689
transit   38   SUBC  0

8

passage   42   SUBC  0
du   43   ARTC  1972
faisceau   44   SAV   0
direct   45   SAV   0
au   46   ARTC  1860
faisceau   47   SAV   0

9

centre   50   SUBC  0
de   51   PREP  689
transit   52   SUBC  0

10

faire   57   VBIF  254

11

manuelle   61   SAV   0

12

horloge   67   SAV   0

13

autocommutateur   71   SUBC  0

14

services   72   SUBC  2702
manuels   73   SAV   0
obtenus   74   VBPA  21
par   75   PREP  696
le   76   ARTD  612
pr(fixe   77   SAV   2237

15

abonn(s   80   SAV   0

16

communication   83   SAV   0

17

d(part   86   SUBC  2002
d'   87   PREP  689
un   88   ARTI  613
autocommutateur   89   SUBC  0

18

tax(es   106   SAV   0
par   107   PREP  696
ticket   108   SUBC  0

19

autocommutateurs   110   SAV   0

20

pr(occuper   116   VBIF  1476
de   117   PREP  689
la   118   ARTD  612
taxation   119   SUBC  0
des   120   ARTC  1972
communications   121   SAV   0
demand(es   122   VBPA  1246
par   123   PREP  696
voie   124   SUBC  2431

21

signaux   127   SUBC  0
de   128   PREP  689
ligne   129   SUBC  2798
utilis(s   130   VBPA  2683

22

code   135   SUBC  3260
de   136   PREP  689
signalisation   137   SUBC  0
pour   138   PREP  697
l'   139   ARTD  612
acc)s   140   SUBC  0
@   141   PREP  685
l'   142   ARTD  612
interurbain   143   SAV   0
manuel   144   SAV   0

23

services   146   SUBC  2702
obtenus   147   VBPA  21
par   148   PREP  696
le   149   ARTD  612
pr(fixe   150   SAV   2237

24

appels   153   SUBC  0
@   154   PREP  685
destination   155   SUBC  0
du   156   ARTC  1972
service   157   SUBC  2702
des   158   ARTC  1972
renseignements   159   SUBC  3236
t(l(phoniques   160   SAV   0

25

service   165   SUBC  2702
desservi   166   VBPA  91
par   167   PREP  696
le   168   ARTD  612
pr(fixe   169   SAV   2237

26

pouvoir   172   VBIF  116

27

faisceau   177   SAV   0
direct   178   SAV   0
```

```
28

interm(diaire   181   SUBC  0
d'   182   PREP  689
un   183   ARTI  613
centre   184   SUBC  0
de   185   PREP  689
transit   186   SUBC  0

29

service   189   SUBC  2702

30

local   191   SAV   0

31

lignes   194   SUBC  2798
d'   195   PREP  689
abonn(s   196   SUBC  0

32

acheminements   202   SAV   0

33

pouvoir   204   VBIF  116
coexister   205   VBIF  0

34

autocommutateur   209   SAV   0

35

cas   211   SUBC  3255

36

faisceaux   219   SAV   0
distincts   220   SAV   0

37

cas   225   SUBC  3255
d'   226   PREP  689
un   227   ARTI  613
acheminement   228   SUBC  0
par   229   PREP  696
faisceau   230   SAV   0
direct   231   SAV   0

38

utiliser   236   VBIF  2683
de   237   PREP  689
signaux   238   SUBC  0
d'   239   PREP  689
enregistreurs   240   SUBC  0

39

signaux   242   SUBC  0
de   243   PREP  689
ligne   244   SUBC  2798
utilis(s   245   VBPA  2683

40

code   250   SUBC  3260
de   251   PREP  689
signalisation   252   SUBC  0
pour   253   PREP  697
l'   254   ARTD  612
acc)s   255   SUBC  0
au   256   ARTC  1860
service   257   SUBC  2702
des   258   ARTC  1972
renseignements   259   SUBC  3236

41

interm(diaire   268   SUBC  0
d'   269   PREP  689
un   270   ARTI  613
centre   271   SUBC  0
de   272   PREP  689
transit   273   SUBC  0

42

signalisation   275   SAV   0
utilis(e   276   VBPA  2683

43

code   280   SUBC  3260

44

ligne   285   SUBC  2798
```

ch13

chap 13

essais des lignes et des circuits

introduction

Ce chapitre pr{sente dans une premi)re partie les besoins que doit satisfaire le syst)me pour les essais de lignes et de postes d'abonn{s,et dans une seconde partie les besoins relatifs aux essais de circuits interurbains et de jonctions urbaines.
En ce qui concerne la maintenance des circuits interurbains,les besoins exprim{s repr{sentent le maximum de ce qui peut ^etre demand{ @ un autocommutateur de rattachement d'abonn{s,les besoins r{els {tant fonction du nombre de circuits interurbains raccord{ @ chaque autocommutateur.En ce qui concerne le dispositif SEQUIN dont le raccordement n'est pas actuellement demand{ pour ce type d'autocommutateur,le syst)me doit :

disposer de joncteurs de liaison,

pr{voir le raccordement ult{rieur @ un dispositif de ce type dont le mode de raccordement serait adapt{ aux autocommutateurs @ commande par calculateur.

essai des lignes @ partir du poste

Il doit ^etre fourni un ou plusieurs dispositifs qui r{alisent les m^emes fonctions que le dispositif d'essai rapide des lignes d'abonn{s (DERAL) (1 dispositif au moins pour 5.000 abonn{s) d{crit au tome 2,chapitre 3.2.
Le num{ro d'appels du DERAL doit ^etre unique pour la zone desservie par un m^eme centre de commutation.

essai des lignes par op{rateur

description des dispositifs d'essais des lignes et des postes d'abonn{s

robots d'essai des lignes d'abonn{s (RELA)

Ils permettent les essais de la ligne d'abonn{ et du poste @ cadran.

D'autre part,ils doivent permettre de d'effectuer des tests syst{matiques lanc{s par l'unit{ de commande du syst)me.De ce fait,les diff{rents tests command{s @ partir de la position d'op{ratrice doivent ^etre supervis{s par l'unit{ de commande.
Plusieurs robots doivent pouvoir travailler simultan{ment dans le m^eme central.

Le RELA doit disposer de 3 liaisons :

une interface normalis{e,appel{e interface de maintenance,permettant la commande de 32 appareils,
2 raccordements au r{seau de connexion,ces raccordements pouvant ^etre r{alis{s :
soit directement,

extraction des Groupes Conceptuels sans regroupement

1

lignes 5 SUBC 2798

2

circuits 8 SUBC 0

3

introduction 9 SUBC 0

4

satisfaire 21 VBIF 259

5

essais 26 SUBC 0
de 27 PREP 689
lignes 28 SUBC 2798

6

postes 31 SUBC 0
d' 32 PREP 689
abonn(s 33 SUBC 0

7

essais 43 SUBC 0

8

maintenance 56 SUBC 3242

9

maximum 65 SUBC 887

10

autocommutateur 74 SUBC 0

11

abonn(s 78 SUBC 0

12

nombre 85 SUBC 3219

13

autocommutateur 112 SUBC 0

14

disposer 116 VBIF 1274
de 117 PREP 689
joncteurs 118 SUBC 0
de 119 PREP 689
liaison 120 SUBC 0

15

pr(voir 121 VBIF 106

16

dispositif 127 SUBC 0

17

calculateur 143 SUBC 0

18

lignes 146 SUBC 2798
@ 147 PREP 685
partir 148 VBIF 51

19

poste 150 SUBC 0

20

dispositif 165 SUBC 0

21

lignes 170 SUBC 2798

22

num(ro 188 SUBC 0
d' 189 PREP 689
appels 190 SUBC 0

23

commutation 205 SUBC 0

24

lignes 208 SUBC 2798
par 209 PREP 696
op(rateur 210 SUBC 0

25

dispositifs 213 SUBC 0
d' 214 PREP 689
essais 215 SUBC 0
des 216 ARTC 1972
lignes 217 SUBC 2798

26

postes 220 SUBC 0
d' 221 PREP 689
abonn(s 222 SUBC 0

27

essai 225 SUBC 0
des 226 ARTC 1972
lignes 227 SUBC 2798

28

essais 234 SUBC 0
de 235 PREP 689
la 236 ARTD 612
ligne 237 SUBC 2798

29

poste 242 SUBC 0
@ 243 PREP 685
cadran 244 SUBC 0

30

effectuer 251 VBIF 1297

31

syst)me 262 SUBC 0

32

partir 271 VBIF 51
de 272 PREP 689
la 273 ARTD 612
position 274 SUBC 0

33

pouvoir 288 VBIF 116
travailler 289 VBIF 1604

34

disposer 297 VBIF 1274

35

maintenance 307 SUBC 3242

36

r(seau 317 SUBC 0
de 318 PREP 689
connexion 319 SUBC 0

extraction des Groupes Conceptuels avec regroupement

```
1

chap   1   SAV   0

2

lignes   5   SUBC  2798

3

circuits   8   SUBC  0

4

introduction   9   SUBC  0

5

partie   16   SAV   2185

6

besoins   18   SAV   0

7

satisfaire   21   VBIF  259

8

essais   26   SUBC  0
de   27   PREP  689
lignes   28   SUBC  2798

9

postes   31   SUBC  0
d'   32   PREP  689
abonn{s   33   SUBC  0

10

partie   38   SAV   2185

11

besoins   40   SAV   0
relatifs   41   ADJQ  3197
aux   42   ARTC  1860
essais   43   SUBC  0
de   44   PREP  689
circuits   45   SAV   0
interurbains   46   SAV   0

12

jonctions   49   SAV   0
urbaines   50   SAV   0

13

maintenance   56   SUBC  3242
des   57   ARTC  1972
circuits   58   SAV   0
interurbains   59   SAV   0

14

besoins   61   SAV   0
exprim{s   62   VBPA  1336

15

maximum   65   SUBC  887

16

autocommutateur   74   SUBC  0

17

abonn{s   78   SUBC  0

18

besoins   80   SAV   0
r{els   81   SAV   0

19

fonction   83   SAV   0
du   84   ARTC  1972
nombre   85   SUBC  3219
de   86   PREP  689
circuits   87   SAV   0
interurbains   88   SAV   0
raccord{   89   SAV   0

20

autocommutateur   92   SAV   0

21

dispositif   98   SAV   0

22

autocommutateur   112   SUBC  0

23

syst}me   114   SAV   0

24

disposer   116   VBIF  1274
de   117   PREP  689
joncteurs   118   SUBC  0
de   119   PREP  689
liaison   120   SUBC  0

25

pr{voir   121   VBIF  106

26

dispositif   127   SUBC  0

27

adapt{   137   SAV   0

28

calculateur   143   SUBC  0

29

lignes   146   SUBC  2798
@   147   PREP  685
partir   148   VBIF  51

30

poste   150   SUBC  0

31

dispositifs   158   SAV   0

32

dispositif   165   SUBC  0
d'   166   PREP  689
essai   167   SAV   0

33

lignes   170   SUBC  2798
d'   171   PREP  689
abonn{s   172   SAV   0

34

dispositif   175   SAV   0

35

abonn{s   180   SAV   0

36

tome   183   SAV   0
```

```
37

num(ro   188   SUBC  0
d'   189   PREP  689
appels   190   SUBC  0

38

zone   198   SAV   0
desservie   199   VBPA  91

39

commutation   205   SUBC  0

40

lignes   208   SUBC  2798
par   209   PREP  696
op(rateur   210   SUBC  0

41

description   211   SAV   0
des   212   ARTC  1972
dispositifs   213   SUBC  0
d'   214   PREP  689
essais   215   SUBC  0
des   216   ARTC  1972
lignes   217   SUBC  2798

42

postes   220   SUBC  0
d'   221   PREP  689
abonn(s   222   SUBC  0

43

robots   223   SAV   0
d'   224   PREP  689
essai   225   SUBC  0
des   226   ARTC  1972
lignes   227   SUBC  2798
d'   228   PREP  689
abonn(s   229   SAV   0

44

essais   234   SUBC  0
de   235   PREP  689
la   236   ARTD  612
ligne   237   SUBC  2798
d'   238   PREP  689
abonn(   239   SAV   0

45

poste   242   SUBC  0
@   243   PREP  685
cadran   244   SUBC  0

46

effectuer   251   VBIF  1297

47

tests   253   SAV   0
syst(matiques   254   SAV   0
lanc(s   255   VBPA  1787

48

syst)me   262   SUBC  0

49

tests   268   SAV   0
command(s   269   VBPA  1199
@   270   PREP  685
partir   271   VBIF  51

50

position   274   SUBC  0
d'   275   PREP  689
op(ratrice   276   SAV   0

51

robots   286   SAV   0

52

pouvoir   288   VBIF  116
travailler   289   VBIF  1604

53

central   293   SAV   0

54

disposer   297   VBIF  1274

55

maintenance   307   SUBC  3242

56

appareils   313   SAV   0

57

raccordements   315   SAV   0
au   316   ARTC  1860
r(seau   317   SUBC  0
de   318   PREP  689
connexion   319   SUBC  0

58

raccòrdements   321   SAV   0
```

AUTORISATION DE SOUTENANCE

DOCTORAT 3ème CYCLE, DOCTORAT INGENIEUR,
DOCTORAT DE L'UNIVERSITE JOSEPH FOURIER - GRENOBLE 1

Vu les dispositions de l'Arrêté du 16 avril 1974,

Vu les dispositions de l'Arrêté du 5 juillet 1984,

Vu les rapports de M ...Jacques...COURTIN.....

M ...Patrice...POGNAN.....

M ....Patrick...PALMER.....est autorisé(e)
à présenter une thèse en vue de l'obtention du Doctorat de.....
l'Université Joseph FOURIER-GRENOBLE 1.....

Grenoble, le .23.JUIL.1990.....

Le Président de l'Université
Joseph Fourier - Grenoble 1

A. NEMOZ

UNIVERSITE JOSEPH FOURIER
GRENOBLE I