BISKRI I., JOUIS C., LE PRIOL F., DESCLES J-P. MEUNIER J-G., MUSTAFA W.(1997)
Outil d'aide à la fouille documentaire : approche hybride numérique linguistique
*FRACTAL 97*, 10-12 décembre 1997, Besançon, France, pp 35-44
*Conférence internationale : Linguistique et Informatique : Théories et Outils pour le traitement Automatique des Langues.*

# Outil d'aide à la fouille documentaire : approche hybride numérique linguistique

**Ismail Biskri (3), Christophe Jouis (1,2), Florence Le Priol (2), Jean-Pierre Descles (2), Jean-Guy Meunier (3), Widad Mustafa (1)**
(1) IDIST/CREDO, Université Charles de Gaulle - Lille 3
(2) LALIC - CAMS, Université de la Sorbonne - Paris IV
(3) LANCI, Université du Québec à Montréal - UQAM

**ABSTRACT :**

Computer-aided Knowledge extraction from extensive textual data is generally based on two methods, the statistical and the linguistic. These two methods were, till recently, considered by specialists as divergent. In this article we endeavor to show that they are rather complementary. If we focus on the two methods we can say that the first one tends towards processing extensive data but the results are rather rough while the second one, based on semantic analyses yields better results. It allows moreover a structured knowledge representation. The idea is to combine these two strategies in an integrated production process in order to have better results. Two systems can be considered : CONTERM, a connectionist-based model and SEEK, a semantic-based one. The integration of SEEK to CONTERM is considered in the framework of FRANCIL (Francophone Language Engineering Network). This research project is conducted in collaboration between LANCI, University of Quebec in Montréal, CAMS, Paris IV and CREDO/IDIST, Lille III.

**KEYWORDS :** development of linguistic databases, automatic document systems.

## 1. INTRODUCTION

Le volume de documents sur support électronique ne cesse de prendre de l'ampleur. Il est devenu presque impossible de repérer rapidement à la main de l'information pertinente compte tenu de l'hétérogénéité des textes. Un certain nombre de systèmes informatiques d'analyse semi-automatique de documents textuels (pour l'aide à la construction de terminologies ou de relations sémantiques entre termes, l'aide à l'acquisition ou la modélisation des connaissances, à l'indexation ou au filtrage de documents, etc.) commencent à être opérationnels. Mais ils ont aussi certaines limites liées aux méthodes utilisées. Il existe deux grandes catégories de méthodologies qui peuvent être combinées :

(1) les méthodes " linguistiques " pures (fondées sur des analyses morphologiques, lexicales, syntaxiques, et/ou sémantiques, etc.) ;
(2) les méthodes numériques pures (statistiques ou connexionnistes).

Les méthodes linguistiques pures ont pour objectif une analyse sémantique fine (représentation structurée des connaissances, s'appuyant sur des connaissances syntaxiques et/ou des primitives sémantiques) mais ne peuvent effectivement fonctionner que sur des corpus de taille limitée, en partie à cause du temps de traitement ou de la nécessité d'une intervention humaine. A l'inverse, les méthodes numériques pures ne fournissent des résultats pertinents qu'avec des documents de taille significative. Toutefois, elles permettent d'absorber de gros corpus mais le résultat obtenu est un filtrage numérique grossier (par exemple classification de l'information par cooccurrences). Aussi, une idée intéressante consiste à combiner ces deux méthodes pour obtenir un système plus efficace.

Dans cet article, nous présentons une expérience d'intégration dans une même chaîne de traitement de deux stratégies complémentaires :

(1) une analyse connexioniste (le système CONTERM, conçu par le LANCI) avec
(2) une analyse sémantique linguistique (système SEEK, conçu par l'équipe LALIC et l'IDIST).

Cette expérience est réalisée grâce à une Action de Recherche Partagée (ARP) financée par le réseau FRANCIL[1].

[1] Par ailleurs, ces deux systèmes sont soumis à une évaluation dans le cadre d'une Action de Recherche Concertée (ARC A3) soutenue par l'AUPELF-UREF (Association des Universités Partiellement ou Entièrement de Langue Française) : " Évaluation des systèmes de construction automatique de terminologie et de relations sémantiques entre termes à partir de corpus ". voir (Mustafa & Jouis 1996, Jouis 1996). Ces deux systèmes sont donc en particulier évalués sur des corpus identiques.

Dans la deuxième et la troisième partie nous présentons successivement les principes, les fonctionnalités et les limites actuelles des deux systèmes. Dans la quatrième partie, nous montrons comment est envisagée et expérimentée l'interconnexion de ces deux systèmes.

## 2. LE SYSTEME CONTERM

### 2.1. PRINCIPES

L'approche de CONTERM (pour " Connexionnisme et Terminologie ") vise à l'application d'une méthode connexionniste au problème de l'extraction de connaissances terminologiques à partir de textes pleins. Elle consiste à effectuer un filtrage numérique sur des gros corpus hétérogènes[2] sans utiliser de connaissances préalables sur le ou les domaines traités[3]. Ce filtrage a pour objectif de classifier et de structurer un corpus en des classes de termes qui serviront d'indices de régularités d'associations lexicales que le terminologue peut utiliser comme tremplin pour approfondir les étapes ultérieures d'interprétation, de construction de réseaux sémantiques, et finalement d'élaboration de ses fiches terminologiques.

### 2.2. FONCTIONNALITES

Une plate-forme réalisée au LANCI, en l'occurrence, la plate-forme ALADIN (Meunier, J.G.& Seffah, A., 1995) permet d'exécuter une chaîne de traitement qui réalise un tel filtrage. La chaîne présente les étapes suivantes : elle commence par une gestion du document, suivent alors un filtrage du lexique et une description morphologique (lemmatisation). Le filtrage du lexique consiste à éliminer du lexique les mots fonctionnels non porteur de sens ainsi que d autres mots que l'utilisateur ne jugera pas pertinents. La lemmatisation consiste à remplacer chaque mot par son équivalent canonique. (e.g. aimerions --> AIMER ). Ce processus se justifie par le fait que les déclinaisons propres à la grammaire ou à la syntaxe d'une langue n'affectent en rien le contenu sémantique prédicatif des termes. De la même façon, remplacer un mot décliné (soit dans sa forme verbale, adverbiale, adjectivale, pronominale ou autres) par sa forme nominale n'a aucun impact significatif sur le contenu sémantique principal de ce dernier. Ces dimensions morphologiques touchent surtout des modalités telles que : le genre, l'aspect, le temps, etc.

Puis une transformation est opérée pour obtenir une représentation matricielle du texte. Cette transformation est encore effectuée par des modules de CONTERM explicitement dédiés à cette fin. On produit ainsi un fichier indiquant pour tout lemme choisi sa fréquence dans chaque segment du texte. Suit ensuite un post-traitement pour construire une matrice dans un format acceptable par les réseaux de neurones. Vient ensuite une extraction classificatoire par réseaux de neurones FUZZYART.

Le réseau neuronal génère une matrice de résultats qui représentent la classification trouvée. Chaque ligne (ou vecteur) de cette matrice est constituée d'éléments binaires ordonnés. La ligne indique pour chaque terme du lexique original s'il fait ou non partie du prototype de la classe. Ainsi est créé un "prototype" pour chacune des classes identifiées. On dira alors que la classe no. X est "caractérisée" par la présence d'un certain nombre de termes. Autrement dit, chaque classe identifie quels sont les termes qui se retrouvent dans les segments de textes qui présentent, selon le réseau de neurones une certaine similarité. Ainsi, les classes créées sont caractérisées, arbitrairement, par les termes qui sont présents également dans tous les segments du texte qui ont été "classifiés" dans une même classe.

Les résultats du réseau de neurones se présentent donc (avant interprétation) sous la forme d'une séquence de classes que l'on dira "caractérisées" par des termes donnés et incluant un certain nombre de segments.

Le système CONTERM permet ainsi de construire des classes de fragments qui entretiennent entre eux une ressemblance en ce qu'il partagent des unités linguistiques communes. A partir de ces classes, il est alors possibles de construire des réseaux d'associations entre unités linguistiques. Ce réseau est interprétable selon la thèse associative classique qui dit que si deux termes se retrouvent ensemble dans un même contexte c'est que leur contenu sémantique ou conceptuel est associé. Ce type de résultat produit par CONTERM permet alors au terminologue d'identifier le réseau des lexèmes spécifique au texte et donc les connaissances particulières à ce texte.

---

[2] c'est-à-dire constitué de textes traitant de sujets différents ou connexes. Un même terme présent dans différentes zones textuelles peut alors avoir des significations différentes suivant le contexte dans lequel il est inséré.

[3] comme c'est parfois le cas dans les systèmes d'Intelligence Artificielle classiques où il faut parfois introduire des connaissances préalables au domaine (lexique, représentations sémantiques, " frames ", " scenario ") pour obtenir des représentations.

Autrement dit, il s'agit de repérer divers fragments de textes dans lesquels des termes coocurrent de manière régulière. Les ensembles de termes constituent alors des classes hiérarchisées. Les sous-textes identifiées sont considérés comme des zones contextuelles dans lesquelles les termes sont supposés entretenir des liens de nature sémantique.

Prenons un exemple. Dans l'expérimentation d'un premier corpus de tests de l'ARC A3 [4], nous avons le terme *rapport* qui apparaît dans les classes *28, 35, 39, 40, 54, etc.*
La classe 28 est composée des termes : *choix, connaissance, document , façon, fiction, personnage, rapport, savoir et travail.*
La classe 35 contient : *autres, connaissances, doute, formes, image, processus, production et rapport*
La classe 39 est constituée par : *autres, élèves, enseignant, ensemble, genre, jeunesse, rapport, roman.*
La classe 40 est définie par : *écrit, écriture, élémentaires, jeunesse , monde, problème, rapport, réel, scolaire, situation.*
Dans la classe 54, nous avons : *auteurs, autres, discours, jeunesse, lecture, mode, question, rapport, rôle, temps, vie.*

Cette liste montre que le terme *rapport* est utilisé dans deux ensembles de contextes relativement différents.
Un premier pointe vers le concept de *rapport* comme *document* où est déposé de l'information (classe 28).
Un deuxième pointe vers le concept des *liens entre des individus et autres chose* (classes 39, 40, 54).
Enfin la classe 35 n'est pas clairement intégrable dans un des deux sens précédents.
On voit que dans le texte, ces deux significations sont les deux seules possibles. Pour un terminologue, le terme rapport dans la Revue Spirale #15 n'est donc pas employé dans les sens suivants : d'une proportion, c'est-à-dire d'un rapport logique, d'un rapport financier , d'une maison de rapport, d'une communication, d'une perspective, etc.

### 2.3. LIMITES ACTUELLES

Les résultats produit doivent ainsi être interprétés. En particulier, il est nécessaire d'expliciter les relations sémantiques que les termes entretiennent entre eux dans une même classe, d'où l'idée d'intégrer un outil linguistique d'aide à l'interprétation : le système SEEK. Dans la suite, nous présentons ce système, puis nous montrons comment l'intégrer à CONTERM.

## 3. LE SYSTEME SEEK

### 3.1 PRINCIPES

L'objectif du système SEEK (pour " Système Expert d'Exploration (K)contextuelle ") est une analyse sémantique de textes en français pour l'aide à l'acquisition et à la modélisation d'un domaine de connaissances. Les textes analysés sont essentiellement des documents techniques ou scientifiques de description d'objets complexes en vue de leur modélisation. Il détecte les relations statiques entre les objets du domaine (hiérarchies classe/sous-classe, attribut/valeur, instances de classe, relations entre un objet et ses parties, identifications, comparaisons, incompatibilités, etc.). Ces relations sont insérées dans un système organisé de significations. Le modèle linguistique sous-jacent est celui de la Grammaire Applicative et Cognitive (GAC), équipe LALIC, (Desclés 1990) : les significations des unités linguistiques sont représentées à l'aide de primitives sémantiques (indépendantes d'un domaine particulier) qui définissent un système organisé de significations.

Le système produit une représentation visuelle sous forme de graphes conceptuels de Sowa enrichis. Des liens hypertextes permettent de retrouver, pour chaque relation identifiée, la partie du texte ayant servi à sa construction. SEEK est un module d'analyse sémantique qui ne nécessite ni dictionnaire, ni d'analyse syntaxique complète pour fonctionner.

### 3.2 METHODOLOGIE

Le système s'appuie sur l'exploration contextuelle, une méthode linguistique et informatique qui permet en particulier de repérer des relations sémantiques entre les termes dans un texte. Cette méthode n'utilise pas de connaissances préalables concernant le monde externe mais un savoir linguistique qui tend à être indépendant d'un domaine de compétence particulier. L'exploration contextuelle aboutit à la prise de décisions fondées sur le repérage des indices (marqueurs linguistiques de relations sémantiques) co-présents dans le texte. Un système d'exploration contextuelle se ramène à un ensemble de règles déclaratives qui

[4]Il s'agit de la revue Spirale, une revue semestrielle de recherche en éducation. Le corpus est composé de 18 numéros de 200 pages environ.

expriment un savoir décisionnel interprétatif. Les règles (d'exploration contextuelle) se représentent sous la forme **SI** <conditions> **ALORS** <actions> ou <conclusions>. Les conditions des règles expriment la co-présence ou non d'unités linguistiques pertinentes dans le contexte. Ces indices touchent plusieurs composantes simultanément : morphologique, syntaxique, lexicale. Les conclusions de l'ensemble des règles permettent de construire progressivement des représentations sémantiques.

La Grammaire Applicative et Cognitive articule plusieurs niveaux de représentations, et en particulier un niveau cognitif où l'on analyse les significations des unités linguistiques sous forme de représentations sémantico-cognitives afin de construire les représentations des connaissances associées à un texte. Ainsi, pour construire des réseaux de concepts et de relations, SEEK s'appuie sur des primitives sémantiques et sur un ensemble de règles d'exploration contextuelle.

### 3.2.1. Les primitives sémantiques

Les primitives sémantiques définissent un système organisé de significations. Nous distinguons :

1. un système de types sémantiques constitués à partir de types élémentaires pour les unités linguistiques (entités individualisables, entités booléennes, entités massives, classes collectives, classes distributives, les lieux, etc.);
2. des opérateurs formateurs de types complexes à partir des types élémentaires ( listes, n-uplets, types fonctionnels, etc. );
3. des relations statiques fondamentales (que nous décrivons dans la suite) ;
4. des relations évolutives (mouvement, changement d'état, conservation d'un mouvement, itération, variation d'intensité, contraintes, causes...)[5].

Les relations statiques sont des relations binaires. Elles permettent de décrire des états (situations statiques) du domaine d'expertise. Les situations statiques restent stables pendant un certain intervalle temporel où ni début, ni fin ne sont envisagés. Nous distinguons plus d'une vingtaine de relations, en particulier : (1) les identifications entre deux entités (*R, c'est le rayon de l'objectif);* (2) les incompatibilités entre entités (*Les chromosomes qui diffèrent suivant le sexe sont appelés gonosomes, par opposition au reste des chromosomes, appelés autosomes)* ; (3) les mesures: (*Le camion pèse 15 tonnes) ;* (4) les cardinalités (*Les français sont 56 millions)* ; (5) les comparaisons (*Le chiffre d'affaire 94 de la société l'emporte sur celui de 95*); (6) les inclusions entre classes distributives (*La famille des particules élémentaires est disjointe en 3 catégories : les protons, les neutrons et les électrons*) ; (7) les relations d'appartenance d'une entité individuelle à une classe distributive (*Citons par exemple le filtre passe-bas.*) ; (8) les localisations d'une entité par rapport à un lieu : intérieur, extérieur, frontière, fermeture (*Le boîtier contient la carte électronique*); (9) les relations partie/tout (ingrédience) entre classes collectives (*L'ammoniac résulte de la décomposition des matières organique)*; (10) la possession (*Le premier secrétaire du parti a un grand bureau à sa disposition)*; (11) les attributions (*A chaque voiture on associe un numéro de série, un numéro d'immatriculation*) ; etc.

La sémantique de chaque relation statique correspond à des propriétés intrinsèques :

1. type fonctionnel (type sémantique des arguments de la relation) ;
2. propriétés algébriques (réflexivité, symétrie, transitivité, etc.) ;
3. propriétés d'agencement (combinaison) avec les autres relations dans un même contexte (c'est-à-dire dans une situation statique donnée). Les relations statiques s'insèrent dans un système de significations des relations de repérage entre entités[6].

### 3.2.2. Les règles d'exploration contextuelle

D'autre part, le système SEEK est composé de règles d'exploration contextuelle organisées sous la forme d'un système de prise de décision. Ces règles permettent d'associer à des cooccurrences d'unités linguistiques la représentation sémantique adéquate en fonction du contexte. Une analyse sémantique de textes faisant appel à l'exploration contextuelle se présente sous la forme d'un système à base de connaissances dont le moteur d'inférences utilise une base de règles associées à une base de données de marqueurs. Ces règles et ces marqueurs expriment un savoir linguistique. Cette approche ne nécessite pas une analyse syntaxique

[5] Dans la suite, nous ne détaillerons qu'une partie des primitives de la GAC : les relations statiques car ce sont ces primitives qui sont pour le moment utilisées dans SEEK. Pour plus de détails sur les types sémantiques et les relations évolutives, nous renvoyons par exemple à (Abraham 1995) et (Descles 1990).

[6] Le repérage, noté REP (ou "ε"), est un schéma général de relation : une entité X (une entité repérée) est repérée par rapport à Y (une entité repère). Le repérage se spécifie suivant les propriétés algébriques qui lui sont attribuées axiomatiquement en divers relateurs. Sur ce point, voir (Desclés 1987).

complète du texte. Il s'agit d'imiter la stratégie d'un lecteur qui a peu de connaissances sur le domaine traité par le texte. Parmi ces règles, nous avons par exemple :

SOIT x1, x2, x3, x4, x5 des unités linguistiques ; P une proposition
SI x1 un marqueur de la liste **auxinc**
et x2 est un marqueur de la liste **verbeinc5**
et x3 est un numéral ou fait partie de la liste **inc6**
et x4 est un marqueur de la liste **inc7** ou **inc1**
et x5 est un marqueur de la liste **inc8** ou **app2** ou l'un des symboles de ponctuation **2-points** ou **parenthese-ouvrante**
et x1 x2 x3 x4 x5 se suivent (pas forcément immédiatement) dans la même proposition P
ALORS Proposer la relation **d'inclusion** ou d'**appartenance** dans la proposition P

Les listes de marqueurs linguistiques mentionnées dans la règle d'exploration contextuelle sont les suivantes :

**auxinc** = {on, nous, il est de usage de, il est usuel de ...}
**verbeinc5** = {distinguer, reconnaître, différencier, discriminer, isoler, séparer, distribuer, discerner, singulariser, particulariser, remarquer ...}
**inc6** = {différent, plusieurs, de nombreux, divers, beaucoup de, un certain nombre, un grand nombre ...}
**inc1** = {sous-classe, classe, sous- caste, caste, sous- catégorie, catégorie, sous- groupe, groupe, sous- division, division, sous- espèce, espèce, sorte, race, sous- ensemble, ensemble, sous- variété, variété, type, sous- type, prototype, archétype, modèle, pattern, sous- famille, famille, sous- genre, genre, collection, concept, partie, partition ...}
**inc7** = {forme, état, disposition, consistance, apparence, aspect, modalité, configuration, conformation, style, structure, nature ...}
**inc8** = {dont, tel que, telle que, telles que ...}
**app2** = {notamment, notablement, par exemple, en particulier, comme, nommément, entre autres, particulièrement, spécialement, surtout ...}

Par exemple, sur l'énoncé ci-dessous, la règle pourra se déclencher pour proposer la relation d'inclusion ou d'appartenance :

> *En fait, on peut considérer que nous distinguons deux familles de dispositifs : les informations et les commandes.*

La base de connaissances de SEEK concernant la détection des relations statiques est composée d'une base de données de marqueurs statiques (quelques 3300 marqueurs classés dans 240 listes) et de 220 règles d'exploration contextuelle[7].

### 3.3. LIMITES ACTUELLES

Au départ, SEEK avait été conçu dans un but d'acquisition des connaissances à partir de textes. Il a été initialement testé sur un grand nombre de petits textes d'expertises dans des domaines techniques différents[8]. Mais, pour des textes plus longs (par exemple les corpus utilisés pour l'évaluation dans l'ARC A3 : 200 à 300 pages), nous rencontrons deux types de points à améliorer :

1. Le phénomène du *bruit*, c'est-à-dire que certaines relations détectées ne sont pas pertinentes pour deux raisons. D'une part, certaines règles se déclenchent trop souvent (une analyse linguistique plus fine dans ce but est en cours ainsi que l'ajout d'un lemmatiseur). D'autre part, sur un corpus de taille importante et hétérogène, certains termes apparaissent avec des significations différentes (polysémie). De ce fait, certains graphes de termes et de relations contiennent des relations qui sont incohérentes entre elles dans

[7] Pour la version actuelle de SEEK. Pour des raisons de place, nous ne donnons ici qu'un extrait du contenu des listes présentées. Ces données sont en cours d'évolution pour tenir compte des évaluations dans le cadre de l'ARC A3 précédemment mentionné en introduction. Ces modifications nécessitent une analyse linguistique fine puisque la méthode sera d'autant plus efficace que les listes d'indices linguistiques seront nombreuses, organisées et relativement complètes

[8] La taille de ces textes d'expertise, correspondant à des entrevues de durée limitée avec des experts en vue de la modélisation des connaissances pour conception de systèmes à base de connaissances, est de l'ordre de trois à quatre pages. Ces entrevues avaient été réalisées dans le cadre des sociétés EDIAT/CR2A/CGI, IPSé et ALSTHOM. Citons en particulier le projet AGENOR (d'EDIAT), un système de gestion des nomenclatures pour des pièces mécaniques.

un même contexte. Il apparaît donc nécessaire de construire plusieurs graphes pour un même corpus qui correspondent à des points de vue différents[9].

**2.** La représentation sous forme de graphe devient finalement difficile à lire, du fait du nombre rapidement très important de noeuds (termes) et de relations entre les noeuds[10]. Face à ce problème, trois solutions sont envisagées. La première solution consiste à ajouter des algorithmes de représentation partielle du graphe complet (détection de cliques par exemple), mais avec un risque de perte d'information. La deuxième solution consiste à changer de mode de présentation. L'idée actuellement retenue revient à présenter les résultats dans un système *de bases de données relationnelles (tableau TERME-RELATION-TERME)*. Cette deuxième approche a l'avantage de permettre ensuite d'envoyer des requêtes de sélection sur les résultats. Ces requêtes permettent de répondre à des questions du type : Pour un terme donné, quels sont les autres termes en relation ? Pour un type de relation donné, (par exemple la hiérarchies générique/spécifique) quels sont les termes qui entretiennent ce type de relation ? Avec ces requêtes, on arrive ainsi à extraire des informations suivant les objectifs d'utilisation des résultats (configuration) : indexation, recherche d'informations, acquisition des connaissances, etc. On obtient alors un sous-tableau à partir duquel il est toujours possible de reconstituer un graphe (plus petit). La troisième solution est de combiner SEEK et CONTERM.

## 4. INTEGRATION DE CONTERM ET SEEK

Cette intégration se fait en deux étapes successives. Tout d'abord, CONTERM est appliqué sur le gros corpus. Il transmet à SEEK des sous-textes dans lesquels apparaissent de façon régulière des classes d'unités linguistiques[11].Vient ensuite le traitement par SEEK qui s'applique donc sur des sous-textes de taille plus réduite dans lesquels les termes sont utilisés dans un contexte bien déterminé. Ces sous-textes sont considérés comme des zones textuelles dans lesquelles les termes entretiennent des liens de nature sémantique. Il s'agit alors d'utiliser SEEK sur ces zones textuelles afin de déterminer les relations sémantiques que les termes de chaque classe entretiennent entre eux.

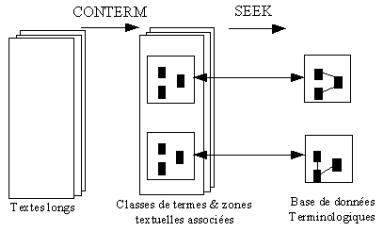


Figure 1 : Intégration de CONTERM et SEEK

Après une expérimentation, nous avons constaté qu'il n'y avait pas forcément de relation directe entre les termes d'une classe donnée issue de CONTERM. Autrement dit, certains termes ne sont reliés à d'autres termes de la classe qu'en passant par des " chemins " de longueur supérieure à un.

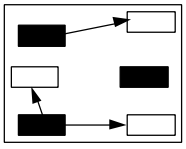

Figure 2 : Phase initiale du traitement par SEEK

Aussi, nous avons mis au point un processus itératif d'application de SEEK dans le but de relier en plusieurs étapes les termes de chaque classe. A l'étape initiale, SEEK n'examine que les énoncés des sous-textes associés à la classe qui contiennent les termes de la classe considérée. Le système tente alors de trouver des relations entre les termes de la classe.

A l'issue de cette première étape, si des relations sont trouvées, elles sont de deux types. D'une part ce sont des relations " directes " : elles relient deux termes de la classe (les termes de la classe étant représenté par des rectangles noirs dans la figure 2). D'autre part, ce sont des relations qui relient un terme de la classe avec

[9] Ceci reste encore vrai si le corpus concerne un même domaine. En effet, des auteurs différents peuvent présenter des théories ou des méthodes divergentes sur le même sujet.
[10] Le problème peut se résumé ainsi : Comment représenter sur un plan un réseau de noeuds et de relations de façon optimale (pas de croisements des liens, espace minimal occupé, regroupement des noeuds, etc...) ?
[11] Cette décompostion peut être répétée de manière récursive de façon à former un treillis de classes.

un terme présent dans le sous-texte mais ne faisant pas partie de la classe (ce deuxième type de terme étant représenté par des rectangles blancs dans la figure 2). Il s'agit alors de réitérer le même processus en ne considérant cette fois ci que les énoncés des sous-textes associés à la classe qui contiennent les termes non présents dans la classe mais déjà en relation.

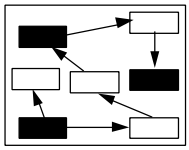

Figure 3 : Itération du processus

Il s'agit ensuite de réitérer ce processus, jusqu'à obtenir des liens sémantiques entre les éléments de la classe jusqu'à saturation de la base de règles de SEEK (plus aucune règles applicables). Le résultat visé est une base de données terminologiques dans laquelle chaque enregistrement est un terme auquel on associe les informations suivantes :

〈 un renvoi vers la (ou les) classes de termes auxquelles appartient le terme ;
〈 pour chaque classe de termes, les zones textuelles du corpus associées (par l'intermédiaire de liens hypertextes) ;
〈 les relations sémantiques que le terme entretient avec les autres termes de la base.

## 5. CONCLUSIONS

L'idée d'associer une analyse linguistique à une analyse numérique est très prometteuse. Elle est également très pertinente, en ce sens qu'elle associe la finesse d'analyse des méthodes linguistiques à la capacité des méthodes numériques de rendre compte de gros corpus. L'ordre stratégique d'appliquer une méthode numérique avant de faire intervenir une méthode linguistique est nécessaire du fait que la méthode numérique est plus adaptée au découpage d'un gros texte et permet à un terminologue de soumettre des segments choisis à l'analyseur linguistique plus fine afin d'en faciliter l'interprétation.

Les résultats encourageant laissent entrevoir des possibilités d'extension des analyses linguistiques de SEEK sur les classes de CONTERM par exploration contextuelle. En particulier, citons la détection des liens sémantiques de types cinématiques ou dynamiques (mouvements d'objets, changements d'états d'objets, relations de causalité, recherche des contextes définitoires d'un terme, etc.), ce qui élargirait les interprétations par le terminologues des niveaux descriptifs statiques du domaine vers des descriptions évolutives (Le Priol 1998).

Enfin un autre modèle hybride est exploré (Biskri, Meunier, 97), il consiste à associer Grammaires Catégorielles et modèles numériques. Ceci laisse entrevoir de nouvelles voies pour le traitement des gros corpus.

## 6. REFERENCES BIBLIOGRAPHIQUES

Abraham M., (1995) *Analyse sémantico-cognitive des verbes de mouvement et d'activité ; Contributions méthodologique à la constitution d'un dictionnaire informatique des verbes*, Thèse de doctorat EHESS, Paris

Biskri, I., (1995). *La Grammaire Catégorielle Combinatoire Applicative dans le cadre de la Grammaire Applicative et Cognitive*, Thèse de Doctorat, EHESS, Paris.

Biskri, I., Desclés, J.-P. & Jouis, C., (1997) " La Grammaire Catégorielle Combinatoire Applicative appliquée au français ", Lexicomatique et Dictionnairiques. Actes des Vème Journée Scientifique du réseau thématique *LTT "Lexicologie Terminologie, Traduction"*, Tunis, 25-27 sept. 97, Beyrouth, Collection " Actualité Scientifique " de l'AUPELF-UREF et F.M.A, à paraître.

Biskri, I., Meunier, J.-G. & Jouis, C., (1997) " Un modèle hybride pour l'extraction des connaissances: le numérique et le linguistique ", Lexicomatique et Dictionnairiques. Actes des Vème Journée Scientifique du réseau *thématique LTT "Lexicologie Terminologie, Traduction"*, Tunis, 25-27 sept. 97, Beyrouth, Collection " Actualité Scientifique " de l'AUPELF-UREF et F.M.A, à paraître.

Bouchaffra, D. & Meunier, J.G. (1995). Markovian Random Field Approach to Information Retrieval. In *ICDAR, IEEE, Computer Society Press. Vol 2*. p. 997-1003

Bourigault, D. (1992), Surface grammatical analysis for the extraction of terminological noun phrases. In *COLING'92: Proceedings of the 15th International Conferences on Computational Linguistics*, August 1992, Nantes, France.

Descles, J.-P. (1990). *Langages applicatifs, langues naturelles et cognition*. Hermès eds., Paris, France.

Descles, J.-P., & Jouis, C., (1993). " L'exploration contextuelle : une méthode linguistique et informatique pour l'analyse automatique de textes ". In *Actes du colloque Informatique et Langues Naturelles (ILN'93)*. Nantes, 2- 3 déc., pp. 339- 350

Desclés, J.-P., (1987). " Réseaux sémantiques : La nature logique et linguistique des relateurs ". In *Langages* n° 87, pp. 55-78

Jouis, C ., Mustafa, W. (1997), " AUPELF Project : Term and Semantic Relation Extraction Tools. Evaluation Paradigms ", In *Proceedings of the Speech and Language Technology (SALT) Club Workshop " Evalutation in Speech and Language Technology "*, University of Sheffied, June 17-18, Sheffield, UK, pp. 106-113

Jouis, C, (1995), SEEK, un logiciel d'acquisition des connaissances utilisant un savoir linguistique sans employer de connaissances sur le monde externe. In *Actes des 6ème Journées Acquisition, Validation, (JAVA 95)*, INRIA, pp. 159--172, Grenoble, Avril 95

Jouis, C, Mustafa-Elhadi, W. (1995). " Conceptual Modeling of Database Schema using linguistic knowledge. Application to Terminological databases " , In *First Workshop on Application of Natural Language to Databases (NLDB'95)*, June 95, Versailles, France. pp. 103-118.

Jouis, C, Mustafa-Elhadi, W. (1996). " Vers un nouvel outil interactif d'aide à la conception de dictionnaires électroniques spécialisés " *Lexicomatique et Dictionnairiques. Actes des IVème Journée Scientifique du réseau thématique "Lexicologie Terminologie, Traduction"*, Lyon, sept. 95, Beyrouth, Collection " Actualité Scientifique " de l'AUPELF-UREF et F.M.A., pp. 255- 266

Jouis, C. & Mustafa, W. (1996), " Approche sémantique par exploration contextuelle pour l'aide à la construction de terminologies. Vers une intégration à une approche statistique ". *Journées de sémantiques Lexicales Brestoises (JSLB'96)*, TELECOM-Bretagne, Brest, 11-13 sept. 1996

Jouis, C. (1996), " Le système SEEK et l'extraction des connaissances statiques dans un texte ". In *Extraction des connaissances et technologies de l'information*, 2ème colloque Franco-québécois, LANCI, Université du Québec à Montréal, 27 sept. 96

Jouis, C., (1993). *Contributions à la conceptualisation et à la modélisation des connaissances à partir d'une analyse linguistique de textes. Réalisation d'un prototype : le système SEEK*, Thèse de doctorat EHESS, Paris

Jouis, C., (1994). " Contextual Approach : SEEK, a linguistic and computational tool for use in knowledge acquisition ". *In Proceeding of the First European Conference " Cognitive Science in Industry "*, 28th - 30th September 1994, Luxembourg, pp. 259-274

Jouis, C., Biskri, I., Descles, J.P., Le Priol, F., Meunier, J.G., Mustafa, W., & Nault, G., (1997), " Vers l'intégration d'une approche sémantique linguistique et d'une approche numérique pour un outil d'aide à la construction de bases terminologiques ". *Actes des Premières Journées Scientifiques et Techniques (JST'97*), FRANCIL, AUPELF-UREF, Avignon, avril 1997, pp. 427-432

Jouis, C., Mustafa Elhadi, W., Biskri, I., Le Priol, F., (1997), " Vers la spécification et l'extension des relations terminologiques : typologie et insertion dans un système de significations des relations ", in *Lexicomatique et Dictionnairiques. Actes des Vème Journée Scientifique du réseau thématique LTT "Lexicologie Terminologie, Traduction"*, Tunis, 25-27 sept. 97, Beyrouth, Collection " Actualité Scientifique " de l'AUPELF-UREF et F.M.A, à paraître.

Béguin, A, Jouis, C., Mustafa, W, (1997) : "Evaluation d'outils d'aide à la construction de terminologie et de relations sémantiques entre termes à partir de corpus", *Actes des Premières Journées Scientifiques et Techniques (JST'97)*, FRANCIL, AUPELF-UREF, Avignon, avril 1997, pp. 419-426.

Le Priol F. (1998), Extraction des relations dynamiques à partir de textes par exploration contextuelle, Thèse en cours, Université Paris-Sorbonne - Paris IV, équipe LaLIC.

Lerat, P. (1988) Terminologie & sémantique descriptive. in *La Banque des mots*, numéro spécial, pp. 11-30.

Meunier, J.G (1996) Théorie cognitive: son impact sur le traitement de l'information textuelle. In V. Riale et D. Fisette : *Penser L'esprit ,Des sciences de la cognition à une philosophie cognitive*. Presses de l'Université de Grenoble. (pp. 289--305)

Meunier, J.G. & Nault, G. (1995), Modèles connexionnistes et traitement de l'information textuelle : le modèle ART de Grossberg. In *Cahiers de recherche du LANCI.95.9*, UQAM

Meunier, J.G., Biskri, I., Nault, G. & Nyongwa, M., (1997), " Exploration de classifieurs connexionnistes pour l'analyse de textes assistée par ordinateur ", in *Lexicomatique et Dictionnairiques. Actes des Vème Journée Scientifique du réseau thématique LTT "Lexicologie Terminologie, Traduction"*, Tunis, 25-27 sept. 97, Beyrouth, Collection " Actualité Scientifique " de l'AUPELF-UREF et F.M.A, à paraître.

Meunier, J.G., Nault, G. & Nyongwa, M. (1996), Extraction dynamique et connexionniste de connaissances terminologiques. In *Extraction des connaissances et technologies de l'information*, 2ème colloque Franco-québécois, LANCI, Université du Québec à Montréal, 27 sept. 96

Mustafa, W., Jouis C. (1997), " Natural Language Processing-based Techniques and their Use in Data Modelling and Information Retrieval ", in *6th International Study Conference on Classification Research, Knowledge Organization for Information Retrieval*, FID/CR, ISKO and University College of London, London, 16-19 june 1997.

Mustafa-Elhadi, W. & Jouis, C. (1996), Evaluating Natural Language Processing Systems as a Tool for Building Terminological Databases. In *Proceedings of the Fourth International ISKO Conference : Knowledge Organization and Change, Advances in Knowledge Organization*, Vol.5, INDEX Verlag, Frankfurt/Main, (pp 346--355)

Mustafa-Elhadi, W. & Jouis, C. (1996), Natural Language Processing-based Systems for Terminological Construction and their Contribution to Information Retrieval". In *Proceedings of the Fourth International Congress on Terminology and Knowledge Engineering (TKE'96)*, Vienna, INDEX Verlag, Frankfurt/Main. (pp 118--130)

Sager, J. C. (1990). A Practical Course in Terminology Processing, John Benjamins Publishing Company, Amsterdam/Philadelphia, 1990.

Salton, G. (1989): Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer. Adisson-Wesley Publishing Company.