Méthode automatique d'annotations sémantiques et indexation de documents textuels pour l'extraction d'objets pédagogiques

Boutheina Smine

Université de Paris-Sorbonne
École Doctorale Concepts et Langages
Equipe STIH
(Sens, Texte, Informatique, Histoire)
Laboratoire LaLIC
(Langues, Logiques, Informatique et Cognition)

Université de Tunis
Institut Supérieur de Gestion
Laboratoire LARODEC
(Recherche Opérationnelle, Décision
et Contrôle de Processus)

Thèse de doctorat

Pour obtenir le grade de Docteur de l' UNIVERSITÉ PARIS IV-SORBONNE
Discipline : "Mathématiques, Informatique Appliquées aux Sciences de l'homme"
Spécialité : Informatique
Cotutelle internationale avec Université de Tunis-Institut Supérieur de Gestion

Présentée et soutenue par :
Boutheina SMINE

Le : **18 Janvier 2014**

Méthode automatique d'annotations sémantiques et indexation de documents textuels pour l'extraction d'objets pédagogiques

Sous la direction de :

Mr. J-P. Desclés	Professeur à l'Université Paris Sorbonne
Mme. R. Faiz	Professeur à L'Université de Carthage (IHEC)

JURY

Mr. J-G Ganascia	Professeur à l'université Pierre et Marie-Curie	Président
Mme. Lamia Belguith	Professeur à l'Université de Sfax (FSEG)	Rapporteur
Mr. Patrice Pognan	Professeur à l'INALCO	Rapporteur
Mr. J-P. Desclés	Professeur à l'Université Paris Sorbonne	Directeur
Mme. R. Faiz	Professeur à L'Université de Carthage (IHEC)	Directeur

Remerciements

Le travail présenté dans cette thèse a été réalisé au sein des deux laboratoires de recherche : LaLIC (Langage, Logique, Informatique et Cognition) à l'université Paris-Sorbonne, et LARODEC (Laboratoire de Recherche Opérationnelle, de Décision et de Contrôle de Processus) à l'ISG de Tunis. Il est l'aboutissement d'un travail en cotutelle, qui a pu être mené à bien grâce à l'interaction et au soutien de nombreuses personnes.

Je tiens tout particulièrment à exprimer toute ma gratitude et mes vifs remerciements à Monsieur Jean-Pierre Desclés, Professeur à l'Uinversité Paris Sorbonne, et Mme Rim Faiz, Professeur à l'Institut des Hautes Etudes Commerciales de Carthage qui m'ont accueillie dans leurs laboratoires en m'offrant un cadre de travail et de réflexion favorables, qui m'ont intégrée dans les projets des laboratoires LaLIC et LARODEC me donnant ainsi l'occasion de rencontrer des chercheurs d'horizons variés.

Je remercie Madame Lamia Hadriche Belguith et Monsieur Patrice Pognan de rapporter ce travail et dont les remarques enrichissantes ont contribué à améliorer ce mémoire. Qu'ils trouvent ici, le témoignage de ma profonde reconnaissance.

Je remercie également ma collègue Madame Asma Bouhafs Hafsia qui, dès le début de ma thèse, n'a cessé de m'orienter et de m'aider dans mes travaux de recherche.

Mes vifs remerciements s'adressent à Monsieur Mohamed Habib Ben Slimen, Maître assistant à l'Institut Supérieur des Langues Appliquées et d'Informatique de Nabeul, pour sa lecture attentive de ce rapport et l'enrichissement qu'il a apporté.

Je souhaite exprimer mon amitié aux membres des deux laboratoires LaLIC et LARODEC de m'avoir accueillie pendant quatre ans en m'accompagant dans mon apprentissage de recherche. Ils ont égayé ma période de thèse par leur compagnie, leurs discussions et leurs sourires.

Je remercie toutes les personnes que j'ai côtoyées durant ces années surtout Monsieur Mohamed Amine Bichiou pour ses précieux conseils et orientations.

Enfin je tiens à remercier vivement mon époux Mohamed pour sa patience, son soutien incessant et son encouragement pendant toutes les années de thèse.

Enfin, je souhaite remercier les membres de ma famille et mes proches surtout ma tante Janette ; je leur suis reconnaissante pour leur soutien constant pendant ces années de travail ; leur encouragement et leur affection ont été déterminants tout au long de la poursuite de mes études.

Dédicaces

A mes adorables parents qui m'ont beaucoup donné

A mon cher époux Mohamed qui m'a beaucoup soutenu

A nos petites filles, les fleurs de notre vie

Table des matières

Liste des figures

Liste des tableaux

Introduction générale

Cette thèse s'inscrit dans le contexte de la recherche d'informations pédagogiques. Notre champ de recherche s'intéresse plus particulièrement à l'extraction d'objets pédagogiques à partir de documents textuels. Ces derniers, appelés aussi " Learning Objects " en anglais, sont devenus centraux sur le terrain des environnements informatiques pour l'apprentissage humain (Pernin, 2003) et font l'objet de nombreux travaux au sein des instances internationales de normalisation. Dans le cadre de notre thèse, nous nous intéressons uniquement aux objets pédagogiques textuels. Ces objets représentent une nécessité pour les enseignants et les apprenants qui les utilisent pour soutenir un processus d'enseignement ou d'apprentissage. Ce besoin n'est pas marginal et quelques exemples peuvent résumer son importance. Prenons le cas d'un enseignant préparant son support de cours dans un domaine bien déterminé, il a nécessairement besoin de consulter des documents s'intéressant à ce domaine pour enrichir ses connaissances, et de là, son cours. Nous citons aussi le cas d'un étudiant préparant son examen, il cherche des exercices sur un concept vu en cours. Toutefois, il existe plusieurs problèmes liés aux outils proposés actuellement à l'utilisateur :

- Le nombre de documents pédagogiques proposés par ces outils est tellement immense, que l'utilisateur n'arrive pas à trouver facilement son besoin.
- La difficulté de choisir rapidement les pages intéressantes au vue des seules informations fournies, les premières lignes par exemple ;
- Le peu d'aide apportée à l'utilisateur (enseignant/apprenant) dans l'expression de son besoin ; même si quelques systèmes proposent des recherches personnalisées comme Google Define.

Ceci est dû, d'une part, à la richesse du langage naturel et d'autre part, à la représentation à base de mots-clés traditionnellement utilisée en Recherche d'Information (RI). Dès lors, quelques défis sont nés de ces problèmes :

- Comment exploiter au mieux les documents pour répondre aux besoins des enseignants et des apprenants ?
- Quelles sont les meilleures méthodes pour permettre à un étudiant d'acquérir rapidement la connaissance recherchée ?
- Est-ce que les outils informatiques existants permettent aux enseignants et apprenants d'assimiler de nouvelles connaissances pour pouvoir constituer des documents pédagogiques ?
- Est-ce que les enseignants et apprenants savent cerner leurs besoins lors de l'utilisation des outils informatiques existants ?

Pour surmonter ces défis, les documents doivent être indexés non en se basant sur des mots-clés comme le font généralement les moteurs de recherche, mais en procédant par une indexation sémantique du contenu des documents, pour passer ensuite à l'extraction automatique de ce contenu.

Les approches d'extraction automatique appliquées jusque-là sont basées essentiellement sur des comptages statistiques de co-occurrences de mots-clefs (Stapley et al, 2000), (Pillet, 2000) ou sur des règles ou automates d'extraction définis manuellement, à base de termes linguistiques. Cependant, le bruit des réponses offertes à l'utilisateur est toujours imminent. Voici un exemple qui illustre une ambiguité dans le mot "cellule" qui peut engendrer un bruit dans un système d'extraction d'informations basé sur des mots clés. En effet, le mot "cellule" peut nous renvoyer à (1) une cellule de prison signifiant " Petite pièce où l'on enferme isolément les détenus dans les prisons ; compartiment d'une voiture cellulaire." (Dictionnaire Larousse) ou encore à (2) une cellule en biologie ayant le sens "Structure microscopique complexe, constitutive de tous les êtres vivants et caractérisée par son pouvoir d'assimilation" (Dictionnaire Larousse).

Pour lutter contre ce bruit, il est nécessaire de mettre en œuvre une méthode d'extraction d'informations plus complexe qui répond à un certain nombre de requêtes formulées par les enseignants et les apprenants pour des besoins de domaines diversifiés et sur différents types de documents.

Le travail que nous présentons dans cette thèse s'inscrit dans cette optique et vise à fournir un système d'extraction d'informations pédagogiques à partir de documents textuels. Afin de faciliter l'exploitation de ces informations, nous avons développé une méthode permettant la mise en valeur d'extraits textuels introduisant des informations

pédagogiques : La définition d'un concept donné, la méthode de résolution d'une équation donnée, des exercices sur une notion donnée, etc.

La question qui se pose alors : Est-ce qu'on peut repérer ces extraits textuels automatiquement et les annoter selon les points de vue de fouille adaptés aux besoins des utilisateurs, dont voici quelques exemples :

- Un enseignant préparant son support de cours a besoin de consulter les définitions du terme "système d'information" pour en choisir celle qui lui convient et l'intégrer dans son support. La définition peut être la suivante :

  "Un système d'information est défini par l'ensemble de moyens humains, matériels et méthodes se rapportant au traitement des différentes formes d'informations rencontrées dans les organisations."

- Un apprenant préparant son examen cherche à pratiquer des exercices sur l'algorithmique. Voici un exemple d'exercice recherché :

  Exercice 1 :
  Ecrire un algorithme qui demande deux nombres a l'utilisateur et l'informe ensuite si leur produit est négatif ou positif (on laisse de cote le cas ou le produit est nul). Attention toutefois : on ne doit pas calculer le produit des deux nombres.

- Un autre enseignant, voulant expliquer une notion donnée pour ses étudiants, cherche à inclure des exemples sur cette notion pour l'enrichissement de son cours.

Ces différents besoins formulés sous forme de scénarios nous amènent à proposer une méthode d'annotation automatique de ces extraits textuels (objets pédagogiques) selon différents points de vue (définition, exercices, exemples, etc.). Ces points de vue permettent de découper l'espace de recherche en sous-espaces correspondant à des approches spécifiques. A chacun de ces points de vue, qui peut être choisi par l'utilisateur, est associé un ensemble de termes, appelés marqueurs linguistiques.
En aval de l'annotation, une indexation par points de vue permet à notre système de se focaliser sur des approches particulières de la notion (définition, exercices, exemples, etc.) qui intéresse l'utilisateur (enseignant/apprenant) ce qui lui offre la possibilité de rechercher les objets pédagogiques dont il a besoin, en diversifiant ses formulations de requêtes selon les différents points de vue.
Notre démarche ne remet donc pas en cause les systèmes éducatifs traditionnels, mais tend à les compléter en combinant la puissance des outils et des ressources informatiques stockées à une méthode linguistique nommée l'exploration contextuelle (Desclés

et al., 1993 ; Desclés, 1997), (Berri, 1996a). C'est une méthode qui permet d'attribuer une valeur sémantique à une entité linguistique en fonction de son contexte en se basant d'abord sur des indices linguistiques du contexte, sans avoir recours à une analyse morphosyntaxique complète préalable.

Dans notre méthode, nous supposons que les textes analysés contiennent un nombre considérable d'objets pédagogiques de différents types (Définition, Exemple, Exercice, etc.), qui sont très utiles pour une bonne annotation, et par la suite, une extraction des informations pertinentes. Cette annotation des objets pédagogiques utilise une méthode qui repose à la fois sur la structure interne des objets pédagogiques, ainsi que sur l'étude du contexte (l'exploration contextuelle).
Comparativement aux méthodes statistiques, classiques en veille, notre méthode devrait permettre de mettre en valeur des informations peu fréquentes, en proposant des informations plus riches.

Cette méthode est par ailleurs utilisable sur différents domaines, puisqu'elle s'appuie sur des connaissances ne dépendant pas du sujet traité. Elle est ainsi adaptée pour l'analyse de corpus traitant de sujets différents. D'ailleurs, dans notre cas, les documents les plus riches en informations pédagogiques sont les documents pédagogiques comme les livres, les supports de cours, les supports de travaux dirigés, les supports de travaux pratiques, les examens, etc. Ces documents ont été principalement récupérés en format électronique. Notre objectif, dans ce contexte, est de réaliser un système d'extraction d'objets pédagogiques, fondé sur l'annotation automatique et sémantique ainsi que sur l'indexation du contenu. En plus de l'extraction des objets pédagogiques répondant à une requête utilisateur, nous proposons à l'utilisateur la constitution de fiches pédagogiques contenant les objets pédagogiques répondant à ses besoins.
Il ne s'agit pas uniquement de proposer une méthode automatisable, mais surtout de réaliser un système pouvant servir les différents utilisateurs qui ont des besoins en informations pédagogiques. Cela a eu une influence dans notre approche du problème : il n'était pas question de proposer une méthode complexe nécessitant plusieurs années de travail pour son élaboration. De même, le système réalisé doit être simple à utiliser pour un utilisateur, n'ayant pas de connaissances particulières en informatique ou en linguistique.

Notre démarche a abouti à la réalisation informatique d'un système, appelé SRIDOP, dont le rôle est l'annotation, l'indexation d'objets pédagogiques en vue de leur extraction en réponse à une requête posée par l'utilisateur. Ces étapes se font d'une manière

complètement automatique. Aucune intervention humaine n'est nécessaire en cours de traitement. Ce système permet aussi de constituer des fiches pédagogiques personnalisables selon le besoin de l'utilisateur. Le travail réalisé dans le cadre de cette thèse n'est pas entièrement achevé. En effet, il n'est qu'une ouverture sur plusieurs domaines, comme le domaine de la pédagogie, de l'enseignement à distance, de l'informatique, etc.

La thèse est constituée de quatre chapitres énoncés ci-dessous.

Le Chapitre 1 **Contexte de l'étude** concerne la définition et la discussion de plusieurs notions liées à notre travail comme la notion d'objet pédagogique. Dans la section 1 de ce chapitre, nous exposerons les différents scénarii d'étude de notre problème. Nous définirons et discuterons respectivement, dans les sections 2, 3, et 4, les notions de document pédagogique, information pédagogique, objet pédagogique. Dans la section 5, nous énoncerons les buts de notre travail et finalement nous présenterons, dans la section 6, quelques approches existantes d'extraction d'informations ; avant d'énoncer les caractéristiques attendues de notre système.

Le Chapitre 2 intitulé **Présentation de quelques approches classiques de l'extraction et de la recherche d'information** débutera par la présentation des étapes de recherche, d'extraction et de traitement de l'information en vue d'une constitution de connaissances (section 1). Par la suite, nous présenterons les principaux acteurs du processus de recherche d'informations, puis nous passerons en revue les principaux modèles qui sont à la base de la majorité des systèmes de recherche d'informations (SRI) existants actuellement (section 2). Les principaux cadres d'évaluation des systèmes expérimentaux seront ensuite décrits (section 3). Enfin, un état de l'art sur l'indexation et la recherche d'objets pédagogiques sera dressé (section 4).

Le Chapitre 3 intitulé **Modèle de l'extraction d'objets pédagogiques à partir de documents** présentera notre méthode pour l'indexation et l'extraction des objets pédagogiques répondant à une requête utilisateur. Dans ce chapitre, nous présenterons tout d'abord nos motivations et le contexte d'utilisation de notre système (section 1), ensuite nous présenterons la méthode d'Exploration Contextuelle et son évolution durant ces dernières années. Par la suite, nous détaillerons notre méthode (section 3) ainsi qu'un exemple d'exécution de cette méthode.

Le Chapitre 4 intitulé **Expérimentations et résultats** aura pour buts (1) de présenter les différents modules de notre système SRIDOP (Système de Recherche d'Informations à partir de Documents pédagogiques) qui implémente notre méthode présentée dans le chapitre 3 (section 1). (2) et d'évaluer ces différents modules (section 2).

Chapitre 1

Contexte de l'étude

Sommaire

1.1 Introduction

La démocratisation de l'informatique dans le monde des particuliers, des entreprises et des administrations a permis de créer des volumes conséquents de documents électroniques rédigés en langue naturelle. Cela a fait naître un besoin d'accès intelligent à cette information textuelle.
La présente étude porte sur l'élaboration d'une méthode et la réalisation d'un système d'extraction d'objets pédagogiques à partir de documents textuels. Cette étude se situe au carrefour de plusieurs domaines : information pédagogique, extraction d'objet pédagogique, le Web sémantique, l'acquisition de connaissances, etc.
Dans ce chapitre, nous exposerons, dans la section 1, différents scénarii d'étude de notre problème. Nous définirons et discuterons respectivement, dans les sections 2, 3, et 4, les notions de document pédagogique, information pédagogique, et d'objet pédagogique. Dans la section 5, nous énonçerons les objectifs de notre travail. Dans la section 6, Nous présenterons, quelques approches existantes d'extraction d'informations, avant de terminer le chapitre par les caractéristiques souhaitées de notre système.

1.2 Scénarii d'étude

Aujourd'hui, partout dans le monde, des milliers d'enseignants intègrent l'utilisation du Web dans leur enseignement, en vue de préparer leurs cours, d'améliorer leur pratique pédagogique, etc. Les apprenants et les étudiants ont également besoin d'accéder à la masse d'informations pédagogiques existantes sur le web, afin d'améliorer leurs connaissances, préparer leurs examens, etc. Grâce aux avancées dans le domaine des technologies de l'informatique, tous ces besoins ont fait que plusieurs scénarii possibles sont créés :

- **Scénario 1** : Un enseignant voulant préparer son support de cours est confronté à plusieurs définitions proposées par différents auteurs. Il a besoin de naviguer à travers ces définitions et choisir celle qui lui convient le mieux. Par exemple : il peut choisir celle qui lui convient parmi les définitions suivantes d'"un système d'information" :

* *Un système d'information (SI) peut être considéré comme un ensemble de flux d'informations, d'opérations qu'ils subissent et de moyens mis en œuvre pour ce faire quelque soit la nature de ces moyens.*

* *Un système d'information est défini par l'ensemble de moyens humains, matériels et méthodes se rapportant au traitement des différentes formes d'informations rencontrées dans les organisations.*

* *Un système d'information peut être constitué de procédures manuelles ou automatisées.*

- **Scénario 2** : Pour expliquer davantage une notion, un enseignant cherchera plusieurs exemples sur cette notion. Par exemple, il souhaite extraire des exemples relatifs aux contraintes dans "le langage de gestion de base de données SQL" comme :

* *Voici quelques exemples de ces contraintes résumé dans une table :*

```
CREATE TABLE T_PATIENT_PTN
(PTN_ID INT NOT NULL PRIMARY KEY,
PTN_NUM_SECU CHAR(13)UNIQUE
,PTN_CLEF_SECU CHAR(2)
CHECK (PTN_CLEF_SECU IS NULL
OR (SUBSTRING(PTN_CLEF_SECU FROM 1 FOR 1)
BETWEEN 0 AND 9) AND
SUBSTRING(PTN_CLEF_SECU FROM 2 FOR 1)
BETWEEN 0 AND 9)),
```

- **Scénario 3** : Un enseignant préparant un nouveau cours sur une notion donnée a besoin de récupérer plusieurs plans de cours sur cette notion pour avoir une idée sur le contenu éventuel de son futur cours. Par exemple, il souhaite trouver ce plan de cours :

Table de matières

- **Scénario 4** : Un étudiant préparant son examen cherche à appliquer plus d'exercices, en dehors de son cours, sur une notion donnée. Par exemple : pour un examen sur "L'algorithmique", l'étudiant pourra extraire du corpus les exercices suivants :

* *Exercice 1 : Ecrire un algorithme qui demande deux nombres a l'utilisateur et l'informe ensuite si leur produit est négatif ou positif (on laisse de cote le cas ou le produit est nul). Attention toutefois : on ne doit pas calculer le produit des deux nombres.*

* *Exercice 2 : Ecrire un algorithme qui demande trois noms a l'utilisateur et l'informe ensuite s'ils sont rangés ou non dans l'ordre alphabétique.*

* *Exercice 3 : Ecrire un algorithme qui demande l'âge d'un enfant a l'utilisateur. Ensuite, il l'informe de sa catégorie :*

  – *"Poussin" de 6 a 7 ans*
  – *"Pupille" de 8 a 9 ans*
  – *"Minime" de 10 a 11 ans*
  – *"Cadet" après 12 ans*

- **Scénario 5** : Un simple utilisateur cherche à savoir les différentes méthodes de réalisation d'une tâche. Par exemple, il veut trouver rapidement une information comme :

La critique de l'information est une méthode qui appartient à l'histoire, c'est une critique historique. C'est une méthode appliquée au passé mais qui peut être utile pour des documents présents.

Ces différents scénarii introduisent des points de vue de fouille (La notion de "point de vue de fouille" sera détaillée plus tard dans ce chapitre) par lesquels est guidé l'utilisateur (enseignant, apprenant). L'hypothèse générale qui sous-tend notre démarche vise à s'inspirer de ce que fait un humain, en particulier notre utilisateur, lorsqu'il souligne certains passages (des segments textuels) ou les annote par des étiquettes sémantiques comme *Définition, Exemple, Exercice,* etc., dans les documents étudiés pour utiliser ultérieurement ces annotations pour constituer par exemple des fiches de lecture, préparer des cours, préparer des examens, etc. Ces segments textuels sont souvent constitutifs de documents pédagogiques (la notion de document pédagogique sera détaillée et discutée dans la section suivante). Nous donnons ci-dessous un exemple d'extrait d'un document pédagogique.

ACCIDENT DE TRAVAIL

DEFINITIONS :

1. D'après le code de la sécurité sociale : **Un accident de travail est considéré comme un accident survenu par le fait ou à l'occasion du travail à toute personne salariée ou travaillant dans le lieu de travail et pendant le trajet d'aller ou de retour entre :**

- *Sa résidence et le lieu de travail*
- *Le lieu de travail et le restaurant, la cantine où d'une manière plus générale, le lieu où le travailleur prend habituellement ses repas.*

2. D'après la cour de cassation : **L'accident de travail est légalement caractérisé par l'occasion violente et soudaine d'une cause extérieure provocant, au cours du travail,une lésion corporelle.**

Nous retrouvons dans cet extrait plusieurs segments textuels :

- Le premier segment textuel suivant reflète *une définition d'un accident de travail.*

D'après le code de la sécurité sociale : **Un accident de travail est considéré comme un accident survenu par le fait ou à l'occasion du travail à toute personne salariée ou travaillant dans le lieu de travail et pendant le trajet d'aller ou de retour entre :**

- *Sa résidence et le lieu de travail.*
- *Le lieu de travail et le restaurant, la cantine où d'une manière plus générale, le lieu où le travailleur prend habituellement ses repas.*

- Le deuxième segment textuel présenté ci-dessous constitue *une caractéristique d'un accident de travail.*

D'après la cour de cassation *: L'accident de travail est légalement caractérisé par l'occasion violente et* ***soudaine*** *d'une cause extérieure provocant, au cours du travail, une lésion corporelle.*

A ce moment-là, ces segments pourront être annotés manuellement ou automatiquement, comme une "Définition" pour le premier segment et une "Caractéristique" pour le deuxième segment. Ces annotations peuvent servir plus tard pour une indexation . Dans ce qui suit, nous détaillerons et discuterons plusieurs notions qui sont au carrefour de notre travail de thèse, à savoir : le document pédagogique, l'information pédagogique et l'objet pédagogique.

Pour la clarté de notre travail, nous précisons que, dans le cadre de notre thèse, nous ne nous intéresserons pas aux domaines de l'e-Learning et de l'enseignement à distance. Par contre, notre travail peut servir ces deux domaines, en offrant une extraction des objets pédagogiques à partir de documents textuels, tout en appliquant l'annotation sémantique de ces objets.

1.3 Document pédagogique

Pour définir le terme "Document pédagogique", il convient de définir tout d'abord le mot "document", ensuite le mot "pédagogie". Le dictionnaire Larousse définit le "Document" comme une pièce écrite servant d'information, de preuve et définit la "pédagogie" comme (1)un ensemble de méthodes utilisées pour éduquer les enfants et les adolescents, (2) une pratique éducative dans un domaine déterminé (méthode d'enseignement) (3) l'aptitude à bien enseigner, sens pédagogique.

D'après Françoise Clerc, la pédagogie est"l'ensemble des savoirs scientifiques et pratiques, des compétences relationnelles et sociales qui sont mobilisées pour concevoir et mettre en œuvre des stratégies d'enseignement"(Marquié,2010).

Un document pédagogique est donc un document dont le contenu pédagogique joue le rôle d'un savoir structuré par l'enseignant sous forme d'un ensemble d'unités de connaissances liées entre elles, dans le but de présenter l'information (connaissances déclaratives), et de fournir un espace d'exploration (activités pédagogiques) ou d'échange (travail collaboratif) (Bousbia *et al.*, 2007).

Le document pédagogique se différencie par rapport à un document classique par le fait qu'il se compose de plusieurs éléments pédagogiques assemblés. Ces éléments sont appelés grains pédagogiques (Flory, 2004)(Cette notion sera approfondie plus loin dans le chapitre)qui fonctionnent comme des lego, pouvant exister seuls et ayant leur propre entité ; ainsi ils peuvent être utilisées dans différents contextes. Un enseignant va constituer son cours en assemblant ces différents grains pédagogiques.

Plusieurs travaux comme (Bertin *et al.*, 2004), (Bodain, 2006), (Cernea *et al.*, 2008) se sont intéressés aux documents pédagogiques. Le travail de (Mille, 2005), par exemple, consiste à proposer des modèles et outils pour l'annotation sémantique des documents pédagogiques. Nous citons aussi le système ProfilDoc (Michel *et al.*, 2002). Les fondements du projet sont basés sur le fait que l'augmentation continuelle de la masse d'information à consulter rend de plus en plus pénible la recherche de l'information pertinente, ceci est d'autant plus vrai lorsque l'on consulte des bases de données en texte

intégral. Le sens d'un texte étant donné, non seulement par son contenu mais aussi par sa structure, l'idée dominante de Profil-doc est que les parties de document auront un usage différencié à priori suivant le besoin de l'utilisateur (Michel *et al.*, 2002). Leurs propos est de vérifier s'il est possible, en considérant les supports pédagogiques comme des documents virtuels personnalisables, de trouver des caractéristiques à la fois fine pour pouvoir personnaliser l'offre de formation aux différents apprenants mais aussi universelles pour être utilisées par le plus grand nombre.

La notion de document introduit certes la notion d'information. Si chaque document a une finalité propre, la manière dont l'information y est représentée renseigne le lecteur sur le type de document et son utilisation (Balpe *et al.*, 1996). D'une manière générale, nous avons toujours besoin d'extraire des informations à partir de documents. L'idée est d'utiliser des méthodes d'extraction automatique basées sur l'annotation pour que le nouveau document puisse être constitué.

1.4 Information pédagogique

Ce concept a été travaillé par la théorie de l'information et la théorie cybernétique. La théorie de l'information, sans précision, est le nom usuel désignant la théorie de l'information de Shannon, qui est une théorie probabiliste permettant de quantifier le contenu moyen en information d'un ensemble de messages, dont le codage informatique satisfait une distribution statistique précise. Ce domaine trouve son origine scientifique avec Claude Shannon qui en est le père fondateur avec son article "A Mathematical Theory of Communications" publié en 1948.
Plusieurs vocables sont souvent utilisés dans le même contexte soit pour définir la même chose, soit pour invoquer des concepts différents : donnée, information, connaissance. C'est pourquoi un effort de clarification doit être entrepris avant toute chose.

D'après le dictionnaire Larousse, "une information est une indication, précision, renseignement, que l'on donne ou que l'on obtient sur quelqu'un ou quelque chose".
Dans le domaine de l'informatique, une information est définie comme un élément de connaissance susceptible d'être représenté à l'aide de conventions pour être conservé, traité ou communiqué.
(Blumentritt *et al.*, 1999), puis (Balmisse, 2002) différencient la donnée de l'information. Pour eux, une donnée est un élément brut livré en dehors de tout contexte. Par exemple, 10 Millions d'Euros est une donnée. Il est impossible de l'interpréter en dehors d'un contexte. Il pourrait s'agir tout aussi bien d'un chiffre d'affaires, d'un résultat d'exploitation, d'un total d'un bilan ou encore d'un prix d'un immeuble. Elle n'a aucune

valeur en soi. Par contre, cette donnée devient une information lorsqu'elle est contextualisée. Si cette valeur de 10 millions d'Euros est avancée alors que la discussion porte sur le résultat d'exploitation d'une entreprise pour l'année 2004, elle prend de la valeur et prend le statut d'information.

La connaissance est un processus dynamique créé à travers une interaction sociale entre individus et organisations. La connaissance est spécifique à un contexte (Paquet, 2006). Les auteurs citent l'exemple suivant : "123 ABC Street" n'est qu'une information qui sans contexte ne signifie rien alors que dire "mon ami David habite au numéro 1234, ABC Street, qui se trouve près de la bibliothèque" constitue une connaissance.

Dans le cadre de notre thèse, nous considérerons les points suivants :

- Les informations doivent être présentées sous un formatage clair et immédiatement interprétable (textuel, visuel, audio,...),
- Les informations doivent être structurées pour être exploitables et pour construire des connaissances,
- Les informations doivent être disponibles au moment où elles sont utiles,
- L'abondance d'informations peut tuer la connaissance
- L'information devient connaissance lorsqu'elle est (1)reliée à d'autres informations (catégorisation des informations) ; (2) pertinente ; (3) validée (sûre, digne de confiance) ; (4) rapidement accessible et exploitable.

Il faut penser donc à une nouvelle façon de construire des connaissances à partir des informations. C'est ce que nous traiterons durant tout notre travail de thèse où nous considérons que les informations annotées, extraites et traitées sont des connaissances.

Du côté de l'information pédagogique, nous pouvons la définir comme étant "*une information destinée à être utilisée ou intégrée dans la mise en œuvre des stratégies d'enseignement orientées vers des classes d'apprenants bien identifiées*". L'information pédagogique est généralement exposée dans un document pédagogique, qui lui-même, se compose de plusieurs éléments pédagogiques assemblés pour offrir aux acteurs (enseignants et étudiants) des outils de communication, d'échange, de partage, de validation des savoirs adaptés à leur besoin et leur rythme.

Une information pédagogique peut être représentée sous forme d'un fragment textuel, au travers d'un document pédagogique, que nous définirons sous l'appellation objet pédagogique. Mais avant de donner notre définition et d'en décrire ses caractéristiques, il convient de récapituler les notations similaires trouvées dans la littérature :

- Dans le projet MacWeb(Nanard *et al.*, 1989), le système manipule des informations interconnectées par des liens sémantiques appelées des "grains d'information". Chaque grain textuel peut être décomposé en plusieurs parties.

- Dans le domaine pédagogique, Tom Murray désigne les différentes parties conceptuelles d'un cours par le terme "thème" (topic) ; lorsque ces thèmes sont instanciés, ils sont appelés présentations (Murray, 1996).

- Henze et al. utilisent les unités d'information sémantique (SIU pour Semantic Information Unit) à chacune desquelles sont associés une unité d'information contenue dans un hyper-livre et des éléments de connaissance (Henze *et al.*, 1999).

- Dans (Delestre, 2000), nous trouvons le terme "Item Didactique" pour qualifier les segments qui vont composer le cours. Ces items didactiques sont regroupés pour former le document pédagogique.

- Dans le projet SEMUSDI, l'auteur parle de "Briques élémentaires". Dans le travail de Ranwez (Chabert-Ranwez, 2002), le nom "Brique d'information" est accordé à une idée qui peut être représentée sous forme d'un document électronique au travers d'un média quelconque.

- (Bousbia *et al.*, 2007) proposent le terme "unités de connaissances" liées entre elles, dans le but de présenter l'information (connaissances déclaratives), et de fournir un espace d'exploration (activités pédagogiques) ou d'échange (travail collaboratif) dans le cadre d'une ressource pédagogique.

Néanmoins, nous avons tenu à mettre en œuvre notre propre terminologie, car elle contient certaines caractéristiques qui nous sont propres. Notre désignation, même si elle adoptée par quelques standards et travaux, se veut particulière à notre cadre de travail de thèse, vu qu'elle possède des caractéristiques qui la distinguent des autres nominations adoptées dans la littérature. Tout ceci sera détaillé dans la section suivante.

1.5 Objet pédagogique

La majorité des travaux sur l'apprentissage et l'enseignement se sont concentrés sur la notion d'objet pédagogique, brique essentielle à partir de laquelle sont constitués les nouveaux documents pédagogiques. Aujourd'hui, le terme d'objet pédagogique est devenu central sur le terrain des environnements informatiques pour l'apprentissage. Pourtant cette notion, si elle semble faire l'objet d'un consensus, demeure encore floue et accepte des définitions différentes.

1.5.1 Définitions

Selon l'IEEE, un objet pédagogique peut être défini comme "toute entité numérique ou non, qui peut être utilisée, réutilisée ou référencée lors d'une formation dispensée à partir d'un support technologique". Cette définition permet de considérer comme objet pédagogique un document imprimé, un cours, un exercice, une étude de cas, une présentation, mais également une salle de cours, un rétroprojecteur, etc.

D'après (Paquette, 2004), un objet pédagogique peut comprendre les matériels, les outils, les services, les personnes et les événements. Il peut être de taille et de nature différentes, tels des textes, des documents audiovisuels, des didacticiels, des présentations ou simulations multimédias, etc.

D'autres chercheurs (Catteau, 2008) considèrent les objets pédagogiques comme informations élémentaires dans un document pédagogique. Ces objets représentent de petites unités d'apprentissage autonomes en ligne. Ils sont suffisamment petits pour être intégrés à une activité pédagogique, une leçon, un module ou un cours.

D'après (Wiley, 2000), un objet pédagogique est une ressource numérique qui peut être réutilisée pour soutenir l'apprentissage.

De plus, d'après (Pernin, 2003), un objet pédagogique est une entité numérique ou non, abstraite ou concrète, qui peut être utilisé, réutilisée ou référencée lors d'une formation. Il existe trois principales classes d'objets pédagogiques :

- Les unités d'apprentissage qui permettent de structurer la formation et de l'organiser dans l'espace et dans le temps ;
- Les activités pédagogiques qui définissent les modalités précises d'acquisition, de validation, de communication d'une ou de plusieurs connaissances ;
- Les ressources pédagogiques, physiques ou numériques, nécessaires à la réalisation des activités.

Notre définition d'un objet pédagogique se situe dans l'intersection de la plus part de ces définitions, tout en possédant ses propres traits, à savoir : "*Un objet pédagogique est un segment textuel annoté destiné à transmettre, exploiter une information pédagogique*".

L'application de l'approche par objets dans le domaine des composants pédagogiques a contribué à la volonté d'en décrire précisément les caractéristiques et les services afin d'en assurer le partage et la réutilisation. Il faut noter que tout ceci peut évoluer lorsqu'il existe un langage commun pour les objets pédagogiques pour pouvoir communiquer. Dans la section suivante, nous présenterons la part des objets pédagogiques dans la standardisation.

1.5.2 Objet pédagogique : Les Standards

L'objectif principal de cette thèse n'est pas d'étudier dans le détail les caractéristiques d'un objet pédagogique. Néanmoins, cette section reprend quelques aspects qui nous paraissent utiles pour aider le lecteur à comprendre notre démarche.

1.5.2.1 LOM (Learning Object Metadata) : Métadonnées relatives aux objets pédagogiques

Dans le cadre du projet ARIADNE, les acteurs du projet ont mis au point une norme Learning Object Metadata (LOM) permettant de caractériser des items didactiques. Dans le cadre du standard IEEE, le LOM (version 1.0) considère qu'un objet pédagogique est "toute entité, sur un support numérique ou non (informatique), pouvant être utilisée pour l'apprentissage, l'enseignement ou la formation". Il se limite à l'ensemble minimal de caractéristiques indispensables pour gérer les objets pédagogiques. Ces caractéristiques sont présentées dans le tableau suivant(cf. Tab.1.1) (Pernin, 2003) :

Catégorie	**Elément**	**Valeurs possibles**
Général	Niveau d'agrégation	Média, leçon, cours, curriculum
Informations techniques	Format	type MIME
	Taille	Exprimée en KO
	Localisation	URL par exemple
	Exigences Techniques	Type technologie Nom (PC-Dos, MS-Windows, MacOS, Unis, Netscape, Exporer), etc.
	Durée des sons, des vidéos, des animations	
Informations pédagogiques	Type d'interactivité	(active, présentation, mixte, indéfini)
	Type d'apprentissage	Exercice, simulation, questionnaire, figure , graphe, diapositive, tableau, texte, examen, expérience, problème, autocontrôle. . .)
	Niveau d'interactivité	Très basse, basse, moyenne, haute, très haute
	Densité sémantique par rapport à la taille ou à la durée	Très basse, basse, moyenne, haute, très haute
	Destinataire	Enseignant, auteur, apprenant, gestionnaire
	Contexte d'utilisation	Primaire, secondaire, cycle universitaire, etc.
	Age ciblé	
	Difficulté vis-à-vis du public ciblé	Très facile, facile, moyen, difficile, très dificile
	Temps moyen d'utilisation	
Relations	Nature de la relation vis-à-vis de l'autre ressource	estPartieDe, estComposéDe, estVersionDe, estBaséSur
	Ressource liée	Identifiant de la ressource, Description de la ressource, etc.

Tableau 1.1: Les caractéristiques des objets pédagogique.

1.5.2.2 SCORM (Sharable Content Object Reference Model) : Modèle de référence pour les objets de contenu partageable

Le consortium ADL (Advanced Distributed Learning) issu d'une initiative du Départemnt de Défense américain, se donne pour objectifs de promouvoir l'utilisation de l'apprentissage basé sur les technologies et le web en particulier, de fournir un modèle de référence permettant de garantir la qualité des contenus en termes de réutilisabilité, accessibilité, pérennité, interopérabilité et de fournir une base solide pour des investissements dans le domaine.

Une de ses principales actions consiste dans l'élaboration de SCORM dont la version

1.2 a été publiée en novembre 2001. SCORM se propose donc de définir les différents types de composants nécessaires à la mise en place d'une solution de formation à partir d'éléments réutilisables. Elle distingue trois niveaux :

- La ressource numérique élémentaire constitue la brique élémentaire : il peut s'agir d'un document simple mais également de tout ensemble d'informations pouvant être délivré vers un client Web ;
- Un objet de contenu partageable : c'est un ensemble cohérent de ressources numériques élémentaires ;
- Un agrégat de contenu est un ensemble de ressources pédagogiques structuré de façon cohérente au sein d'une entité de plus haut niveau, telle qu'un cours, un chapitre, un module, etc. les ressources pédagogiques peuvent être aussi bien des ressources numériques élémentaires que des objets de contenu partageables ;

Pour chaque niveau de composant, SCORM propose de définir un sous-ensemble de métadonnées issues du LOM et établit une table de correspondance pour chacun des niveaux, en indiquant la nature (obligatoire, optionnelle ou " en attente") de chacun des éléments (et leurs sous-éléments) définis du LOM. Comme le souligne (Duval, 1999), les standards permettant la réutilisation et le partage des ressources dans le domaine des outils pédagogiques ne sont pas encore nés.

1.5.2.3 Utilité d'un objet pédagogique

En se basant sur la définition de (Wiley, 2000), les objets pédagogiques ont les caractéristiques suivantes :

- Numériques - ils peuvent être stockés et accessibles par voie électronique ;
- Sont une ressource - ils contiennent l'ensemble des moyens disponibles pour offrir une formation et des objets d'apprentissage réutilisables ;
- Peuvent être utilisés dans plusieurs contextes à des fins multiples ;
- Doivent contribuer au transfert des connaissances ;
- Sont autonomes - chaque objet peut être pris indépendamment ;
- Peuvent être agrégés - ils peuvent être regroupés dans de plus grandes collections de contenu, y compris les structures de cours traditionnels ;

– Sont interopérables - ils ont la capacité d'échanger des informations et d'utiliser les informations qui ont été échangées ;

– Sont étiquetés avec des métadonnées - chaque objet possède des informations descriptives qui lui permettent d'être facilement détectable sur un apprentissage système de gestion ou d'un système de gestion de contenu.

Toutes ces caractéristiques font, principalement que, l'utilisation des objets pédagogiques dans le contexte d'enseignement ou d'apprentissage permet la personnalisation du parcours de chacun et la réduction du temps de conception et la réutilisation des ressources.

Le concept d'objet pédagogique rassemble aussi un certain nombre d'atouts reconnus à différents niveaux économique, pédagogique ou technique (Pernin,2003) :

- Au niveau économique, au cours des dernières décennies, le développement de produits ou services reposant sur l'assemblage et la réutilisation de composants s'est largement répandu.

- Au niveau pédagogique, le concept objet s'accorde bien avec les notions de formation tout au long de la vie. En effet, de nombreuses recherches visent aujourd'hui à améliorer la qualité de la formation en fournissant à chaque apprenant une solution personnalisée prenant en compte un ensemble de facteurs tels que son niveau initial, ses objectifs, son style d'apprentissage, sa disponibilité, son éloignement, etc. Construire une offre formation revient alors à assembler un ensemble de composants adaptés aux besoins spécifiques de l'apprenant.

- Au niveau technique, les apports de l'approche par objets dans le domaine du génie logiciel sont indiscutables. Depuis plus de dix ans, cette approche s'est généralisée et a été formalisée en particulier au travers de la méthode UML (Unified Modeling Language). L'application de l'approche par objet dans le domaine des composants pédagogiques a contribué à la volonté d'en décrire précisément les caractéristiques et les services afin d'en assurer le partage et la réutilisation.

1.5.2.4 Différents niveaux de granularité

Un objet pédagogique est un objet modulaire que l'on peut combiner selon des niveaux de granularité formelle ou de complexité d'un niveau très bas (une définition s'effectue souvent en une phrase) à très élevé (Un programme d'études s'expose avec un document contenant des figures, des tableaux,etc.).

Ainsi nous différencions l'objet pédagogique élémentaire qui est un élément de base de la ressource pédagogique (objet textuel, objet animé, etc.)de l'objet pédagogique composite qui représente une collection cohérente d'objets pédagogiques élémentaires (leçons, modules, cours, sites Web, etc.). C'est cette logique d'agrégation des objets pédagogiques qui impose que chaque objet peut être "retrouvable, réutilisable, indexable". Les objets pédagogiques peuvent avoir non seulement une composante "contenu" (correspondant à des documents) mais aussi une composante "processus", conférant au système la capacité à réagir aux initiatives des usagers, voire à guider ces derniers (comme c'est le cas pour des didacticiels).

Plusieurs travaux se sont intéressés aux des objets pédagogiques. Parmi ces travaux, nous citons (1) le travail de (Lee *et al.*, 2008) qui présente un algorithme d'extension des requêtes utilisateurs à partir d'une ontologie pour l'extraction des objets pédagogiques, (2) le travail de (Meyer *et al.*, 2007) pour la catégorisation des objets pédagogiques en se basant sur le corpus Wikipédia.Cependant, la problématique reste la même : les objets pédagogiques sont encore peu "accessible". Ils sont souvent stockés dans un document ou dans une présentation qui a une forme statique qui n'accepte pas d'être réutilisée et qui n'est pas donc adaptable aux besoins des utilisateurs. Dans la plupart des cas, les parties spécifiques sont assemblées manuellement en utilisant des actions de copier/coller. Cependant il est possible de réutiliser les objets pédagogiques d'une manière beaucoup plus flexible et plus autonome. Ceci représente l'un des buts de notre travail qui seront détaillés dans la section suivante.

1.6 Objectifs de notre travail

Aujourd'hui encore, il semble encore difficile de trouver l'objet pédagogique adéquat à des besoins d'un utilisateur donné. Il apparaît de plus en plus souhaitable de recourir à des systèmes d'extraction d'objets pédagogiques en particulier, dans divers documents, mais pas uniquement dans des documents pédagogiques. Tous ces besoins de masse ont fait que nous avons ressenti le besoin de développer un outil d'aide pour tous ces utilisateurs. Afin de faciliter l'exploitation d'un tel outil dans un contexte d'apprentissage, nous avons développé une méthodologie générale indépendante du domaine basée sur des notions pédagogiques, telles que la Définition, l'exercice, l'exemple, la méthode, etc. Par exemple, l'extrait suivant est associé au type d'objet "Définition" :

> *Le processus unifié est un processus de développement logiciel : il regroupe les activités à mener pour transformer les besoins d'un utilisateur en système logiciel.*

Les systèmes d'extraction d'information à venir se doivent de répondre à des besoins plus précis que les systèmes existants pour satisfaire au mieux les utilisateurs. Lorsque

l'intérêt de l'utilisateur porte sur une donnée factuelle, l'information pertinente ne peut être apportée que par des systèmes dédiés à ce type de tâche. En effet, répondre à une question telle que "Quelle méthode utiliser pour résoudre l'équation x*y/z ?" requiert une analyse en profondeur des documents sélectionnés afin d'en extraire l'information pertinente. L'interrogation peut porter sur n'importe quel domaine ou relever d'un domaine de spécialité.

Les buts généraux de notre travail, au-delà du développement d'une application d'extraction d'informations pédagogiques, sont les suivants :

- Aider les utilisateurs (apprenants, enseignants, étudiants, etc.) à naviguer entre les différents objets pédagogiques pertinents par rapport à leur besoin ;
- Créer des fiches pédagogiques personnalisées pour venir en aide aux apprenants qui veulent approfondir des cours et aux enseignants voulant construire des programmes de cours ;
- Aider les utilisateurs à rechercher leurs objets pédagogiques à partir de corpus (document pédagogique ou autre document illustratif, comme les journaux, les articles scientifiques, etc.) ;
- Assister les étudiants dans la réalisation de leurs travaux (Exposés, Travaux à la maison, etc.) ;
- Aider les étudiants à préparer leurs examens en leur facilitant la recherche des exercices ;
- ...

Ces objectifs répondent aux scénarii présentés précédemment (Section 1). Nous tenons à répondre à ces objectifs en tenant compte de plusieurs facteurs actuels qui en justifient l'importance :

- La massification des étudiants nombreux venant d'horizons différents : notre application devrait offrir une assistance aux étudiants pour les amener à un niveau suffisant ;
- Le trop grand nombre de pages proposées par les outils du Web, en plus de la non-pertinence de classement de ces pages par rapport au besoin de l'utilisateur ;
- L'intégration de notre système à des programmes d'apprentissage à distance ;

- Sachant que les besoins des utilisateurs apprenants diffèrent d'un utilisateur à un autre, notre application offre une meilleure acquisition de connaissances personnalisées et adaptées aux besoins de chaque apprenant.

Nous retrouvons dans la littérature trois types d'applications pour remédier à ces problèmes, qui peuvent être résumés sous le nom "Extraction d'informations pédagogiques" à savoir : Les Systèmes de Question/Réponse traditionnels, les systèmes de repérage des entités nommés et les systèmes d'annotation. Ces applications vont être détaillées dans la section suivante.

1.7 Approches existantes pour l'extraction d'information

Nous citons et détaillons particulièrement trois approches pour l'extraction d'information, qui peuvent être appliquées dans la recherche d'information, à savoir les systèmes de question-réponse, les systèmes de reconnaissance des entités-nommées et les systèmes d'annotation (section 1.6.3).

1.7.1 Systèmes de Question/Réponse

Les systèmes de question/réponse ont pour objectif de fournir une réponse précise à une question posée. Il s'agit d'une tâche plus fine et plus exigeante que la recherche d'information, dans la mesure où il ne s'agit plus de fournir des documents entiers mais des informations spécifiques. Ces systèmes sont censés capables de répondre à des requêtes de la forme : "Quelle est la capitale de l'Inde ?". Le système utilise alors des techniques de traitement automatique des langues afin d'analyser la question et de rechercher une réponse adéquate à l'aide des documents auxquels il a accès.
Contrairement aux moteurs de recherches "classiques", qui proposent une suite de documents classés selon l'estimation de leur intérêt, les systèmes de question/réponse cherchent généralement à reconstruire une réponse en langage naturel et non pas à proposer à l'utilisateur une (longue) liste de documents. Le système START (en anglais) (Katz, 1997) est un exemple de système de question/réponse en ligne. Il propose "New Delhi is the capital of India" comme réponse à la question "What is the capital of the second largest country in Asia ?".

Les principaux systèmes de Question-Réponse se décomposent en trois étapes (Jacquemin *et al.*, 2000) :

Etape 1 : Analyse de la question. En partant d'une question exprimée en langue naturelle, l'analyse de la question permet d'orienter la stratégie de recherche de la réponse

grâce à la détermination des caractéristiques des éléments répondant à la question. L'analyse de la question se déroule classiquement de la façon suivante (Even, 2005) :

Etape 2 : Sélection des documents pertinents. Il est nécessaire de restreindre le champ de recherche de la réponse en sélectionnant un sous-ensemble de textes ou de passages de texte pertinents par rapport à la question. Des moteurs de Recherche d'Information sont utilisés dans ce but : soit de façon classique pour collecter un ensemble de documents, soit adaptés de manière à extraire des passages ou des paragraphes de ces documents. Les requêtes fournies aux moteurs de Recherche d'Information sont élaborées à partir de la liste des mots-clefs issue de la question. Ces requêtes sont souvent étendues avec des synonymes des mots, généralement en se servant de dictionnaires électroniques. Les documents ou passages sont ensuite classés selon leur pertinence.

Etape 3 : Localisation de la réponse. Cette dernière étape consiste à trouver la réponse dans les documents pertinents. Les documents sont découpés en phrases. Ces phrases (phrases-candidates) sont comparées avec la question en se servant des éléments extraits lors de la phase d'analyse de la question. Ensuite une note est attribuée à chaque phrase-candidate. Les phrases les mieux notées sont sélectionnées (phrases-réponses). La réponse est obtenue à partir de ces phrases-réponses. Elle prend généralement la forme d'une phrase ou d'un extrait de texte (au nombre de caractères fixé à 16) contenant la réponse ou des éléments de réponse et est assortie d'une valeur de confiance.

Ces systèmes ont quelques limites. En effet, les anaphores posent problème dans le processus de réponse aux questions, en plus les expressions temporelles compliquent l'analyse des questions posées ainsi que la reconnaissance d'entités nommées ne suffit pas pour répondre efficacement aux questions. Plusieurs pistes sont suivies pour améliorer de tels systèmes. Par exemple, la collecte de données telles que des noms de mesures, des types de questions, etc., facilite l'analyse des documents et des questions.

1.7.2 Systèmes de reconnaissance des entités nommées

La reconnaissance des noms propres ou entités nommées est un problème récurrent dans le traitement automatique de la langue naturelle (TALN), pour l'indexation de textes, la traduction, etc. Cette reconnaissance a été réalisée de façon satisfaisante en extraction d'informations.

D'après (Jacquemin *et al.*,2000), les entités nommées comprennent les organisations (entreprise, administration, musées, etc.), les lieux (villes, régions, fleuves, etc.), les personnes (hommes politiques, vedettes, chefs d'entreprise, etc.) et les numériques (poids,

longueurs, valeurs monétaires, pourcentages, etc.). Les entités nommées peuvent constituer des index très discriminants, et sont souvent des informations demandées. Par exemple, plusieurs entités nommées sont en jeu pour répondre à la question "Quel était le nom du PDG de Peugeot en 1987 ?".

La tâche de reconnaissance des entités nommées et leur extraction consiste à les repérer dans le texte concerné et à leur affecter une étiquette sémantique choisie dans une liste prédéfinie ensuite à les extraire selon le besoin de l'utilisateur.

Nous distinguons trois types de systèmes de repérage d'entités nommées (Bouhafs, 2005) à savoir : (1)Les systèmes fondés sur une base de règles écrites à la main par un concepteur permettant de reconnaître puis d'extraire les entités nommées. (2) Les systèmes à base d'apprentissage fondés sur des techniques d'apprentissage pour apprendre un modèle permettant d'étiqueter les textes de manière adéquate à partir d'un corpus annoté et (3) les systèmes mixtes où un ensemble de règles est généralement appris automatiquement puis révisé par un expert.

Les performances des systèmes de reconnaissance d'entités nommées sont évidemment variables en fonction du type des entités nommées recherchées, de la couverture des dictionnaires et des règles, du style rédactionnel et de la structuration des textes analysés. Mais en général, ils fournissent une bonne précision à défaut d'avoir un bon rappel (Amerdeilh, 2007).

En plus des systèmes de question/réponse et les systèmes de repérage des entités nommées qui présentent quelques inconvénients, nous proposons le concept d'annotation sémantique et automatique pour l'extraction d'informations à partir de textes.

1.7.3 Systèmes d'annotation sémantique et automatique

L'annotation est une information pouvant être liée à diverses entités : un ensemble de documents, un document, un passage, une phrase, un terme, un mot, une image, etc., en vue de donner généralement un sens à l'une de ces entités (Mille, 2005).
D'après (Amardeilh, 2007), l'annotation sémantique consiste à ajouter semi-automatiquement ou automatiquement des métadonnées structurées aux ressources documentaires du web et des intranets des entreprises. C'est une représentation formelle d'un contenu, à l'aide de concepts, de relations et d'instances décrits éventuellement dans une ontologie reliée à la ressource documentaire source. L'annotation sémantique a été déjà utilisée dans plusieurs applications comme par exemple : la classification, l'extraction d'information, le résumé automatique, l'interopérabilité.
Dans cette partie, nous présentons les méthodes appliquées pour l'annotation en vue

d'une extraction des informations : les méthodes statistiques et d'apprentissage, les méthodes d'annotation manuelle et les méthodes de traitement automatique des langues.

1.7.3.1 Méthodes satistiques et d'apprentissage

Les méthodes statistiques sont les premières méthodes qui ont été utilisées pour le traitement des informations contenues dans un texte. Elles sont essentiellement utilisées pour l'étiquetage morpho-syntaxique des textes, l'annotation des textes, la constitution de classes de mots et le calcul de la cooccurrence de couples de mots. Elles reposaient au début sur un simple calcul de cooccurrences sur les termes du texte puis elles ont tenu compte d'autres propriétés statistiques comme la fréquence d'occurrence dans l'ensemble du texte, la fréquence relative et la régularité de la répartition(Bertrand-Gastaldy, 1990). Pour mettre en œuvre les méthodes statistiques, il faut, en premier lieu, identifier le problème à résoudre ; en deuxième lieu, modéliser le problème en faisant apparaître les probabilités de certains événements ; en troisième lieu, construire des estimations des valeurs des probabilités élémentaires précédemment définies à partir des données d'apprentissage ; finalement, les probabilités peuvent être utilisées pour traiter de nouvelles données.

Ces méthodes sont robustes et ne nécessitent pas de connaissances préalables sur le domaine. Nous pensons qu'elles sont très pertinentes pour distinguer des classes d'usage de mots ou de termes dans l'espoir de les organiser en systèmes structurés reflétant une organisation conceptuelle.

1.7.3.2 Méthodes d'annotation manuelle des corpus

Geoffrey Leech (Leech, 1997) défnit l'annotation de corpus comme "la pratique consistant à ajouter des informations linguistiques interprétatives à un corpus de données langagières parlées et/ou écrites. L'annotation décrit également le produit final de ce processus.".Les annotations peuvent être posées soit manuellement par un interprète humain soit de manière automatique par un outil d'analyse. Dans le premier cas, l'interprétation peut refléter une part de la subjectivité de son auteur. Dans le second cas, l'interprétation est entièrement déterminée par les connaissances et l'algorithme incorporés dans l'outil d'analyse.

Nous nous intéressons ici à l'annotation manuelle en tant que tâche exécutée par des agents humains (les annotateurs) (Fort, 2012). D'après(Vernay, 2008), les annotations existantes dans les applications de traitement de texte peuvent être sous la forme d'une

note en marge d'un texte, d'un petit numéro renvoyant plus loin dans le document ou encore d'une note de bas de page.

De la qualité des corpus annotés manuellement dépend plus ou moins directement la qualité des outils créés à partir de ces corpus ou de l'évaluation qui les utilise. Ces corpus annotés doivent donc offrir la meilleure qualité d'annotation possible, ce qui implique de faire intervenir des experts humains dans le processus d'annotation, que ce soit pour annoter directement le corpus ou pour corriger une annotation réalisée automatiquement. Cette phase manuelle est extrêmement fastidieuse et nécessite un travail de longue haleine, de qualité si possible constante. En outre, le coût de développement manuel de ressources linguistiques en général, et de corpus annotés en particulier, est notoirement élevé. En fonction d'un besoin applicatif donné, il faut donc trouver un équilibre entre la qualité attendue, le coût de l'annotation et le volume à annoter (Fort, 2012).

1.7.3.3 Méthodes de traitement automatique des langues

- **Traitement en profondeur**

  Les grands axes du TALN, en s'appuyant sur un découpage méthodologique classique en linguistique, sont les suivants (Jacquemin *et al.*, 2000) :

  - L'analyse morphologique : concerne la détermination des formes des mots et leurs variations de forme. D'un point de vue informatique, un texte est une chaîne de caractères. La première étape de l'analyse d'un texte est la reconnaissance, dans cette chaîne de caractères, d'unités linguistiques de base, les mots, et la mobilisation des informations associées, puisées dans un lexique. Le lexique est la liste des mots de la langue, et associe à chaque mot les informations linguistiques correspondantes : catégorie syntaxique, traits morphosyntaxiques (genre, nombre, etc.), etc.
  - L'analyse syntaxique : s'intéresse à l'agencement des mots et à leurs relations structurelles dans un énoncé. Pour repérer quels mots fonctionnent ensemble dans une phrase, un premier niveau de modélisation consiste à constituer des classes de mots (catégories syntaxiques, parties du discours) possédant un fonctionnement similaire : Nom (N), Verbe (V), Adjectif (A), etc. Des relations entre les mots ou syntagmes sont utiles à l'interprétation des phrases. Les relations grammaticales classiques (sujet-verbe, verbe-objet, verbe-objet-indirect) permettent de représenter la fonction des groupes de mots les uns par rapport aux autres.
  - L'analyse sémantique : se consacre au sens des énoncés. De même que pour la syntaxe, un premier niveau de modélisation consiste à constituer des classes

de mots (catégories sémantiques). Ces classes regroupent des mots dont le sens est proche, ou au minimum des mots qui possèdent certaines propriétés sémantiques communes. Cependant, si en syntaxe on arrive à s'accorder sur des jeux de catégories consensuels, en sémantique aucune classification universelle n'existe. Un mot, même syntaxiquement non ambigu, pourra posséder plusieurs sens.

- L'analyse pragmatique : elle prend en compte le contexte d'énonciation. L'interprétation d'un énoncé dépend de son contexte. Dès que l'on veut traiter une phrase ou plus d'une phrase, cette dimension intervient. Le co-texte désigne le texte qui précède et suit la phrase courante. Au-delà du texte lui-même, les conditions d'énonciations et les connaissances partagées complètent le contexte d'un énoncé. L'interprétation devra donc faire appel à des connaissances sur le monde (scénarios, plans, etc.).

Toutes ces étapes ont fait que le traitement automatique des langues naturelles nécessite énormément de ressources linguistiques, temporelles, humaines pour aboutir à des résultats que le traitement de surface parvient à fournir. En plus, ces étapes sont indépendantes les unes des autres, du fait que, par exemple, une erreur au niveau du découpage des syntagmes ou un mauvais choix lexical de la part de l'analyseur syntaxique sera propagée aux autres niveaux d'analyse et de traitement.

- **Traitement de surface**

  - Automates d'extraction définis manuellement

  Un automate est un mécanisme abstrait capable de reconnaître les phrases d'un langage, c'est-à-dire de déterminer, pour un langage L et une phrase W donnés, si la phrase W appartient ou non au langage L (Audibert, 2003).
  Les automates les plus simples, appelés automates à états finis, ou simplement automates finis, sont des reconnaisseurs pour les langages réguliers. Ils sont formellement définis de la manière suivante :
  Un automate à états finis (AEF) ou automates finis est déni par :

  – Un ensemble fini E d'états, E = e_0, e_1,...,e_n, Ø. A chaque moment dans le processus de reconnaissance, l'automate se trouve dans un état donné. L'état Ø, aussi appelé état trappe représente l'état dans lequel se trouve l'automate en cas de transition illicite.

  – Un état $e_0 \in$ E distingué comme étant l'état initial. C'est l'état dans lequel se trouve l'automate au début du processus de reconnaissance.

- Un ensemble fini F inclut E un autre ensemble fini d'états distingués comme états finaux (ou états terminaux).
- Un alphabet Δ des symboles d'entrée.
- Une fonction de transition $\sum$ qui à tout couple formé d'un état E de et d'un symbole de $\sum$ fait correspondre un ensemble (éventuellement vide) d'états : $\Delta(e_i,a) = \{e_0, e_1,...,e_n, \emptyset\}$

Un mot est dit reconnu par un automate donné s'il existe une séquence de transitions qui permet à cet automate, à partir de l'état initial, d'avancer du premier au dernier symbole du mot et de se trouver alors dans un état final. Une phrase qui n'est pas reconnue est dite rejetée. L'ensemble des phrases reconnues par un automate définit le langage reconnu par cet automate. Malgré leurs atouts, les automates sont connus par leur point faible à savoir le fait que tous les éléments d'un automate sont au même niveau. Ils ont tous la même importance par rapport au contexte dans lequel ils sont appliqués.

- Annotation semi-automatique en l'associant à une ontologie du domaine

D'après (Amardeilh, 2007), l'annotation sémantique consiste à ajouter (semi) automatiquement des métadonnées structurées aux ressources documentaires du web et des intranets des entreprises. C'est une représentation formelle d'un contenu, à l'aide de concepts, de relations et d'instances décrits dans une ontologie reliés à la ressource documentaire source.
Une ontologie représente à la fois cet objet de consensus pour les êtres humains et un objet formel permettant son exploitation par un agent logiciel. Elle se compose de classes, d'attributs et de relations qui définissent et précisent l'utilisation de ces classes et de ces attributs. Elle est décrite dans un langage formel de représentation des connaissances (Exemple : RDF, OWL,...).
Les ontologies fournissent les moyens d'exprimer les concepts d'un domaine (Amardeilh, 2007) :Il s'agit de définir des concepts et de les relier par des relations sémantiques en premier lieu, de réaliser des modèles conceptuels en deuxième lieu, et plus encore dessiner des graphes conceptuels.

- Une ontologie est constituée principalement de concepts (classes, objets abstraits, objets concrets, objets réels...), de relations entre ces concepts et d'attributs relatifs à ces concepts.
- Les instances de concepts font partie de la base de connaissances.
- L'action de définir et d'instancier une base de connaissances est appelée "peuplement d'ontologie", alors que "l'enrichissement d'ontologie" consiste à ajouter des concepts, des attributs ou des relations à l'ontologie existante.

Le processus d'annotation de documents en faisant référence à une ontologie comprend :

- Le repérage des éléments qui correspondent aux concepts de l'ontologie
- L'instanciation est le fait de donner une valeur aux attributs des concepts à l'aide d'informations présentes dans le document
- L'enrichissement est l'ajout d'autres concepts à l'ontologie qui n'étaient pas instanciés avant.

L'annotation semi-automatique des documents en se référant à une ontologie dépend fortement des niveaux de compréhension des annotateurs des informations contenues dans les documents. En plus, ce type d'annotation nécessite la création et le peuplement d'une ontologie de domaine, ce qui est un processus extrêmement coûteux en temps et en ressources. L'ensemble du processus pose également des problèmes en termes de productivité et de qualité. Pour toutes ces raisons, les entreprises cherchent de plus en plus à mettre en place des solutions basées sur l'utilisation d'outils linguistiques pour extraire les informations pertinentes des documents textuels. Ces outils du traitement automatique du langage naturel devront s'intégrer étroitement aux futures applications du Web Sémantique et seront même essentiels au développement, à l'acceptation et à l'utilisation du Web Sémantique.

- Techniques linguistiques et computationnelles d'Exploration Contextuelle

La méthode d'Exploration Contextuelle (Desclés *et al.*, 1991), (Desclés, 1997) (Desclés *et al.*, 2009) est une méthode qui permet d'attribuer une valeur sémantique à une entité linguistique en fonction de son contexte en se basant d'abord sur des indices linguistiques du contexte, sans avoir recours à une analyse morphosyntaxique complète préalable. Cette méthode doit nous permettre d'identifier dans les textes les relations cherchées en s'appuyant sur (Desclés *et al.*, 2007) :

- Des indices linguistiques, appelés aussi marqueurs, qui se décomposent en deux catégories : les indices déclencheurs ou les indicateurs, et les indices complémentaires.
- Un ensemble de règles, dites règles d'exploration contextuelle, qui mettent en relation des indicateurs en présence de certains indices, avec des décisions à prendre. Ces indicateurs se déclenchent pour attribuer à une unité lexicale (mot, phrase, paragraphe, etc.) des étiquettes sémantiques, lorsqu'un certain nombre d'indices (indices déclencheurs ou indices complémentaires) ont été trouvés dans le contexte de l'unité lexicale. Le contexte dans lequel les indices complémentaires sont cherchés est défini par la règle.

– Un ensemble de décisions à prendre tel que l'attribution des étiquettes sémantiques à un segment textuel.

L'exploration contextuelle va être utilisée pour repérer les compléments d'informations, et pour l'analyse sémantique des indicateurs et des indices linguistiques identifiés. L'organisation des ressources et outils linguistiques autour de ces règles d'exploration contextuelle a l'ambition de satisfaire certaines exigences qui sont : Devenir autant que possible indépendants des domaines de connaissance afin de ne pas recommencer à construire des systèmes complexes et volumineux pour chaque domaine et chaque tâche spécifique, et pouvoir toucher l'ensemble des thèmes possibles abordés dans un même document et les relations inattendues entre ces thèmes.

L'action de l'ensemble des règles permet de construire progressivement des représentations sémantiques. L'EC se distingue des méthodes d'apprentissage numériques parce que, face au même problème (par exemple, la désambiguïsation) ces derniers évacuent les nuances de l'analyse linguistique en les remplaçant par une analyse probabiliste basée sur des fréquences et des distances entre mots (Mitchell, 1997). Quant aux méthodes d'apprentissage symboliques, et notamment celles basées sur de patrons, l'EC se distingue par la hiérarchisation qu'elle fait des marqueurs entre indicateurs déclencheurs et indices contextuelles (optionnels, obligatoires, négatifs). La méthode d'EC sera détaillée dans les plus grandes lignes dans le chapitre 3.

1.8 Caractéristiques souhaitées de notre système

Cette thèse propose une méthode d'annotation automatique des objets pédagogiques afin de les extraire à partir de documents. Notre travail voudrait s'inscrire dans un courant qui cherche dans le domaine de l'éducation une aide aux apprenants et aux enseignants dans leurs processus respectifs d'apprentissage et d'enseignement. Notre travail vise à mettre en œuvre une méthode d'annotation des objets pédagogiques basée sur l'analyse linguistique du discours pédagogique, une méthode dont les annotations de nature sémantique puissent être exploitées pour l'aide à l'extraction des objets pédagogiques répondant à une requête utilisateur.

Notre analyse de l'organisation du discours dans les objets pédagogiques s'appuie sur un corpus français de documents pédagogiques. Le but de cette analyse est, dans un premier temps, de repérer des marqueurs linguistiques des objets pédagogiques. Dans

un deuxième temps, ces marqueurs sont organisés sous la forme de règles d'EC qui permettent l'annotation automatique des objets pédagogiques selon leurs types : Définition, Exemple, Exercice, etc.

La réalisation de notre système SRIDOP nécessite alors la mise en œuvre d'une méthode d'extraction d'informations plus complexe qui réponde à un certain nombre de requêtes formulées par l'utilisateur pour le besoin des apprenants et des enseignants, en combinant la puissance des outils et des ressources informatiques stockées à une méthode linguistique d'exploration contextuelle et l'annotation sémantique qu'elle permet.
Cette étude vise la mise au point d'un système opérant sur des documents principalement pédagogiques. Nous aurons l'occasion d'y revenir dans les chapitres qui vont suivre. Nous verrons alors que même dans le cas de domaine assez simple (comme les documents pédagogiques), l'annotation et l'extraction des objets pédagogiques ne sont pas des tâches aisées. Le système développé doit donc être :

- Robuste : avoir des performances correctes sur différents types de corpus et privilégier le rappel par rapport à la précision.
- Adaptable : l'utilisateur doit facilement pouvoir gérer ses ressources linguistiques.
- Multilingue : c'est à dire, l'analyse et la constitution des ressources est transposable d'une langue à une autre, la méthode étant la même et la conceptualisation étant en partie la même.
- Rapide : le système doit être rapide dans son exécution des fonctionnalités offertes par notre système (gestion des ressources linguistiques, annotation, réponse à la requête utilisateur, etc.). Il ne doit pas dépasser un temps limite d'exécution.
- Automatique : les fonctionnalités offertes par notre système sont complètement automatiques.

Notre système doit offrir un certain nombre de fonctionnalités selon le contexte d'utilisation. En effet il doit viser dans un premier temps à répondre aux requêtes posées par l'utilisateur. Ensuite il doit aussi prendre en compte les besoins de l'utilisateur pour réaliser les tâches d'annotation, de constitution de fiches pédagogiques, de gestion des ressources linguistiques, etc.

1.9 Conclusion

Dans le but d'éclaircir le contexte d'étude et la problématique de notre thèse, nous avons cherché dans ce chapitre à spécifier les buts de notre travail en passant par les

différents scénarii d'étude et les approches existantes d'extraction d'informations. Dans le chapitre qui suit, nous entamerons une étude de l'état de l'art sur les modèles de recherche d'informations ainsi que sur les systèmes existants d'indexation des documents pédagogiques.

Chapitre 2

Présentation de quelques approches classiques

Sommaire

2.1 Introduction

L'internet est devenu une source d'information incontournable, notamment pour les étudiants, les enseignants, les apprenants, etc. La quantité de documents pédagogiques accessibles, ainsi que le nombre d'utilisateurs, ne cesse de croître. Retrouver et utiliser une information pédagogique adéquate en réponse à une question posée n'est pas une démarche aisée sur Internet. C'est le rôle des systèmes de recherche d'informations (SRI), de faciliter cette démarche. Ces derniers servent d'interface entre une source contenant des quantités considérables de documents et des utilisateurs cherchant, via des requêtes, des informations susceptibles de se trouver dans cette collection. Elles peuvent être résumées en quatre fonctions, qui sont le stockage de l'information, l'organisation des informations, l'extraction d'informations en réponse à des requêtes utilisateurs et la restitution des informations pertinentes pour ces requêtes. Les questions qui se posent sont les suivantes :

- Est-ce qu'il existe des systèmes de recherche d'informations pédagogiques ?
- Quels sont les systèmes existants qui permettent l'extraction des informations pédagogiques ?
- Est-ce que ces systèmes atteignent la satisfaction des utilisateurs ?

L'objectif de ce chapitre est de dresser le portrait de la recherche d'informations en général, et le portrait de la recherche d'informations pédagogiques en particulier. Nous présenterons, en premier lieu, la relation entre les deux concepts "Extraction d'information" et "Recherche d'information". En deuxième lieu, nous énonçerons quelques concepts liés à la recherche d'informations, ainsi que quelques modèles de recherche d'informations. En troisième lieu, nous illusterons quelques systèmes d'indexation des documents pédagogiques, avant de clôturer le chapitre par une synthèse sur ces différents systèmes et une introduction de notre apport par rapport à ces systèmes.

2.2 Extraction d'informations

La composante "Recherche d'informations" prend alors une part importante dans notre travail, ce qui nous amène à détailler ses différents composants et étapes dans la section suivante.

2.3 Recherche d'informations

2.3.1 Concepts de base de la RI

La Recherche d'Information peut être expliqué comme suit : D'une part, nous avons une personne qui a besoin d'une information. Ce besoin doit être formulé dans une demande convertie en une requête. D'autre part, nous avons des informations stockées dans des collections de documents. Ces dernières sont considérées comme une ressource potentiellement précieuse, mais pour y trouver de l'information, elles ont besoin d'être représenté et ensuite indexés (Morizio, 2006). Le défi est de fournir un bon appariement entre ces deux afin de s'assurer que l'information présentée est pertinente par rapport à la requête initiale. Nous nous intéressons particulièrement à la recherche d'informations à partir de textes, ou recherche documentaire, vue qu'elle est appliquée dans notre thèse.

Un processus de recherche documentaire en texte intégral se décompose en trois étapes principales (Boughanem, 1992 ; Boughanem, 2000) :

2.3.1.1 Indexation des textes sur lesquels portent les interrogations

La représentation des documents est souvent appelée *"Indexation"*. Le processus d'indexation est un des aspects les plus intensifs de recherche d'information. La finalité de l'indexation est de permettre une recherche efficace des informations contenues dans une collection de documents sans avoir à analyser chaque texte de document à chaque interrogation ou recherche. Le processus d'indexation peut inclure le stockage réel du document dans le système, mais souvent les documents sont stockés en partie, par exemple seulement le titre et le résumé, ainsi que des informations sur l'emplacement réel du document.

2.3.1.2 La représentation de la requête posée par l'utilisateur

Les utilisateurs ont un besoin d'information. Le processus de représenter leur besoin d'information est souvent désigné comme le processus de formulation de la requête. La représentation qui en résulte est la requête.

2.3.1.3 La recherche des documents pertinents répondant à la requête utilisateur

La comparaison de la requête par rapport aux représentations des documents est appelée "Appariement Document-Requête". Le processus d'appariement se traduit généralement par une liste de classement des documents. Les utilisateurs peuvent parcourir cette liste des documents à la recherche de l'information dont ils ont besoin.

Les trois processus sont présentés dans la figure suivante (cf. Fig.2.1) où les rectangles, à coins droits, représentent les données et les rectangles, à coin arrondis, représentent les processus. Ces processus s'appuient sur un certain de modèles permettant de sélectionner des informations pertinentes en réponses à une requête utilisateur (Bellot, 2000).

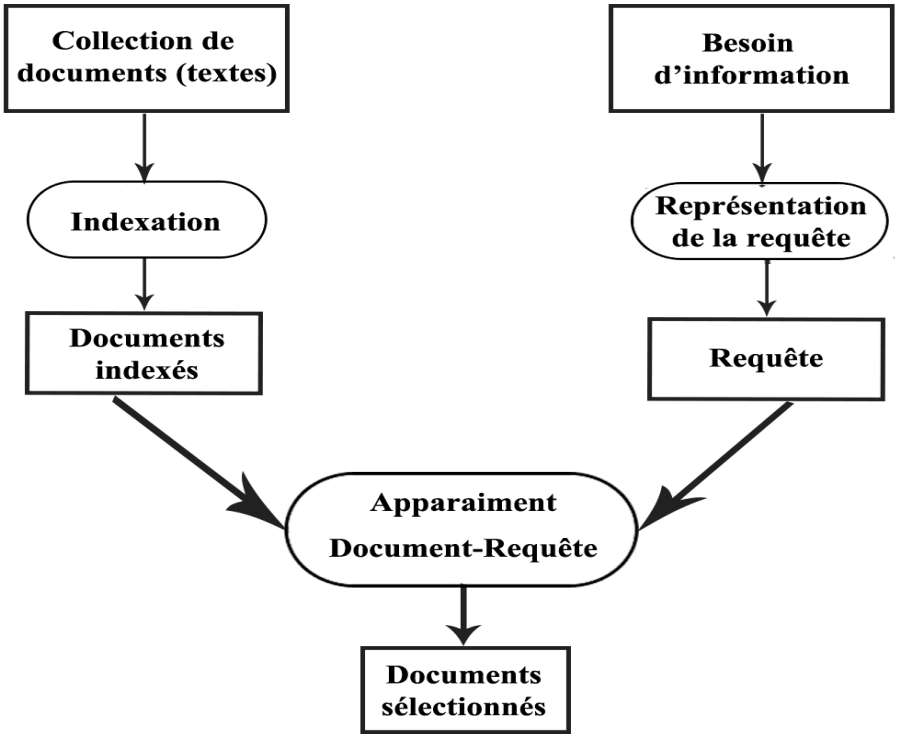


FIGURE 2.1: Processus général de recherche d'information

2.3.2 Les modèles de RI

Un modèle de recherche est un algorithme avec ses structures qui prend une requête et un ensemble de documents et affecte une mesure de similarité entre la requête et chaque document. Cette similitude représente la pertinence par rapport à la requête de l'utilisateur. Les documents sont ensuite classés en fonction de leur similarité vis-à-vis de la requête et présentés à l'utilisateur (Goker *et al.*, 2009). On parle de systèmes de recherche d'informations lorsque l'on désigne une implémentation d'un modèle de recherche d'informations. Par ailleurs, il faut savoir que la plupart des systèmes de recherche d'informations travaillent sur des corpus textuels.

Dans cette section, nous présentons succinctement deux modèles classiques utilisés en recherche documentaires : le modèle booléen, très utilisé par les moteurs de recherche Internet et le modèle probabiliste. Ensuite nous focalisons notre étude sur le modèle vectoriel que nous appliquons dans le cadre de notre thèse.

2.3.2.1 Le modèle Booléen

Le modèle booléen est le premier modèle de recherche d'information et probablement le modèle le plus critiqué. C'est un modèle de recherche documentaire fondé sur l'algèbre de Boole. Dans ce modèle, les requêtes sont exprimées à l'aide d'expressions logiques construites à partir des opérateurs "ou", "et" et "non", et des termes de la requête. La similarité d'un document avec une requête est évaluée selon que les termes du document vérifient ou non l'expression booléenne d'interrogation (cela revient à contrôler l'absence ou la présence des termes de la requête dans le document). Les principaux avantages de ce modèle sont : sa simplicité de mise en œuvre et le formalisme bien défini qui le fonde. Cependant, il propose trop ou pas assez de documents en réponse à une interrogation (Goker *et al.*, 2009). D'une très grande efficacité en termes de temps de réponse, ce modèle et les systèmes qui l'ont implémenté ont eu beaucoup de succès et continuent à être très utilisés. Cependant ses limites sont évidentes puisque la pertinence dans ce modèle est binaire : un document répond ou ne répond pas à une requête. Par exemple le document "t1 et t2" ne répond pas à la requête "t1 et t2 et t3". Le modèle, selon qu'il récupère, ou non, un document, pourrait conduire le système à une bonne, ou mauvaise, prise de décisions. Un avantage du modèle booléen est qu'il offre aux utilisateurs un sentiment de contrôle sur le système (Boughanem, 2006).

La figure suivante (cf. Fig. 2.2) montre un exemple d'extraction des documents répondant à des requêtes bien définies.

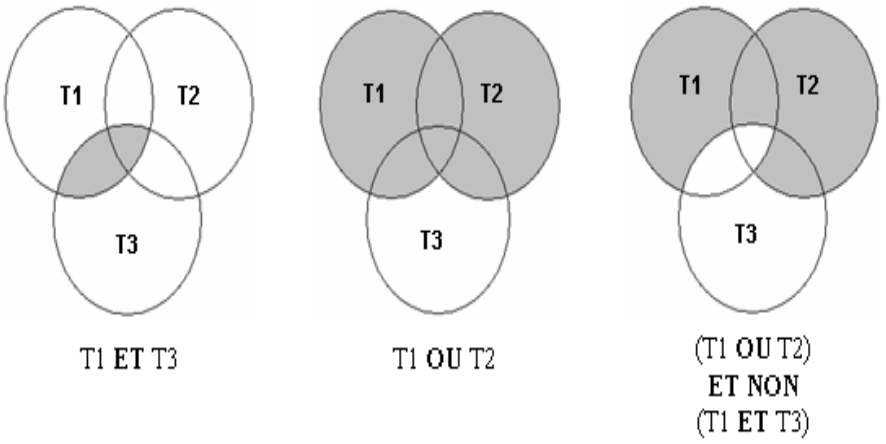


FIGURE 2.2: Exemples d'extraction de requêtes répondant à des requêtes

2.3.2.2 Le modèle Probabiliste

Le modèle probabiliste aborde le problème de la recherche d'information dans un cadre probabiliste. Le premier modèle probabiliste a été proposé par Maron et Kuhns (Maron *et al.*, 1960) au début des années 1960. Le principe de base consiste à présenter les résultats d'un système de recherche d'informations dans un ordre basé sur la probabilité de pertinence d'un document vis-à-vis d'une requête. Robertson (Robertson, 1977) résume ce critère d'ordre par le principe de classement probabiliste, désigné aussi par PRP (Probability Ranking Principle), énoncé comme suit : "Ranking documents in decreasing order of probability of relevance to the user who submitted the query, where probabilities are estimated using all available evidence, produces the best possible effectiveness".

Etant donné une requête utilisateur noté Q et un document D, formellement le modèle PRP peut être traduit de la manière suivante : pour chaque document D et chaque requête Q, quelle est la probabilité que ce document soit pertinent pour cette requête ? Deux possibilités se présentent :

- R, D est pertinent pour Q ;
- $\overline{R}$, D est non pertinent pour Q.

Le modèle probabiliste tente d'estimer la probabilité que le document D appartienne à la classe des documents pertinents (non pertinents). Un document est alors sélectionné

si la probabilité qu'il soit pertinent à Q, notée $P(R|D)$, est supérieure à la probabilité qu'il soit non pertinent Q, notée $P(\overline{R}|D)$. le score d'appariement entre le document D et la requête, noté RSV (Q, D)(Robertson, 1977), est donné par :

$$\text{RSV}=\frac{P(R|D)}{P(\overline{R}|D)}$$

Si l'on applique la règle de Bayes, on a :

$$\text{P(R}|D) = \frac{P(D|R)P(R)}{P(D)} \text{ et P}(\overline{R}|D)=\frac{P(D|\overline{R})P(\overline{R})}{P(D)}$$

En supposant que les documents aient tous la même probabilité d'être sélectionnés et que la sélection d'un document soit indépendante d'un autre, le terme le terme P(D) peut être supprimé. On obtient alors :

$$\text{RSV(Q,D)}=\frac{P(D|R)}{P(D|\overline{R})}\frac{P(R)}{P(\overline{R})}$$

2.3.2.3 Le Modèle Vectoriel

Le modèle vectoriel fait partie des modèles statistiques. L'utilisation des statistiques a pour but, d'une part, de caractériser d'un point de vue quantitatif les termes et les documents, et d'autre part, de mesurer le degré de pertinence d'un document vis-à-vis d'une requête. Le but final est d'arriver à retourner une liste ordonnée de documents selon ce degré. Un autre avantage réside dans l'expression des besoins de l'utilisateur : contrairement au modèle booléen où les termes de la requête doivent être reliés par des connecteurs logiques, l'utilisateur peut ici exprimer son besoin en langage naturel (Boughanem et al., 2008).

Hans Peter Luhn est l'un des premiers à suggérer une approche statistique pour la recherche d'information (Luhn, 1958). Il a suggéré que pour rechercher une collection de documents, l'utilisateur doit d'abord préparer un document qui est semblable aux documents nécessaires. Le mesure de similarité entre la représentation du document préparé et les représentations des documents de la collection est utilisé pour classer les résultats de recherche. La pertinence est ainsi traduite en une similarité vectorielle : *Un document est d'autant plus pertinent à une requête que le vecteur associé est similaire à celui de la requête.*

Une telle définition revient, en fait, à compter le nombre d'éléments qui partagent la requête et la représentation du document. Pour ce faire, considérons la représentation d'un document comme un vecteur :

$$\overrightarrow{dj} = (w_{1,j}, w_{2,j}, ..., w_{n,j})$$

Où $w_{i,j}$ est le poids (0 ou 1) des termes dans le documents, t étant le nombre total de termes de l'index, et considérons la représentation de la requête comme un vecteur

$$\overrightarrow{q} = (w_{1,q}, w_{2,q}, ..., w_{n,q})$$

Avec les mêmes notations. La mesure de similarité la plus simple est alors le produit scalaire :

$$RSV(\overrightarrow{d_j}, \overrightarrow{q}) = \sum_{i=1}^{n} w_{i,j} * w_{i,q}$$

Comme les poids des termes sont binaires, la mesure de similarité mesure le nombre de termes partagés entre le document et la requête. Salton (Salton, 1970) a proposé un modèle basé sur cette mesure de similarité dans son projet SMART (Salton's Magical Atomatic Retriever of Text). Le document (vecteur $\overrightarrow{d}$) et la requête (vecteur $\overrightarrow{q}$) sont représentés dans un espace euclidien de dimension élevée engendré par tous les termes de l'index. La similarité est alors le cosinus de l'angle formé par les deux vecteurs :

$$RSV(\overrightarrow{d}, \overrightarrow{q}) = \frac{\overrightarrow{d}_j * \overrightarrow{q}}{\left|\overrightarrow{d}_j\right| * |\overrightarrow{q}|} = \frac{\sum_{i=1}^{n} w_{i,j} * w_{i,q}}{\sqrt{\sum_{i=1}^{n} w_{i,j}^2} * \sqrt{\sum_{i=1}^{n} w_{i,q}^2}}$$

D'autres fonctions de similarité ont été proposées dans la littérature, parmi lesquelles on peut citer les mesures de Jaccard et Dice.Les documents sont ainsi classés en fonction de la mesure de l'angle qu'ils forment avec le vecteur requête. L'aspect le plus intéressant de cette mesure est l'influence d'un terme isolé sur le score de recherche. Si un terme est présent à la fois dans la requête et le document, il contribue au score. S'il est présent uniquement dans l'un des deux, il diminue le score parce que la requête et le document se correspondent moins.

Plusieurs algorithmes de recherche d'informations ont prouvé leur performance lorsque les vecteurs requête et documents étaient normalisés. L'algorithme d'apprentissage de Rocchio (Rocchio, 1971) en est un exemple.Concernant la pondération des termes, les travaux de Salton (Salton, 1971, 1988) ont montré qu'il ne s'agissait pas d'un problème trivial, mais les pondérations selon tf et idf restent les plus courantes et les plus simples. Les avantages d'une telle pondération sont nombreux : la pondération des termes augmente les performances du système, le modèle permet de renvoyer des documents qui répondent approximativement à la requête, et la fonction d'appariement permet de trier

les documents selon leur degré de similarité avec la requête.

Théoriquement, le modèle vectoriel a l'inconvénient de considérer que les termes de l'index sont tous indépendants. Cependant en pratique, la prise en compte globale de la dépendance des termes peut faire baisser les performances d'un système (puisque les dépendances sont généralement locales) (Buckley, 1994). De nombreuses méthodes d'ordonnancement des résultats ont été comparées au modèle vectoriel, et celui-ci, malgré sa simplicité, est supérieur ou, du moins, aussi bon que les autres alternatives. C'est pour toutes ces raisons qu'aujourd'hui le modèle vectoriel est le plus utilisé dans plusieurs domaines comme en recherche d'information où le modèle vectoriel est appliqué, par exemple, par (Haddad et al., 1996) pour rechercher des informations à partir de documents vidéo et aussi en classification pour classifier leurs données (Claveau et al., 2010), (Cleziou, 2004), (Cleziou et al., 2008).

2.4 Recherche vs. Extraction d'informations en vue d'une construction des connaissances

La tâche principale d'un système de recherche d'information (SRI) est de sélectionner dans une collection de documents les informations qui sont susceptibles de répondre aux besoins des utilisateurs.
Cependant, il ne suffit pas de retrouver une information mais il faut aussi pouvoir la catégoriser pour l'archiver, la synthétiser et la réutiliser en temps utile. La bonne information n'est pas nécessairement celle qui est connue de tous. L'information rare, cachée, émergente est souvent plus utile. La bonne information est nouvelle, pertinente, validée et réutilisable pour construire des connaissances ou entreprendre des actions. Une nouvelle façon de construire des connaissances est alors proposée : Des systèmes d'extraction d'information qui effectuent une analyse de documents bruts afin d'en extraire uniquement des informations précises qui intéresseront l'utilisateur. Par exemple, (Faiz, 2006) propose une méthode d'extraction des évènements à partir des articles. Ces informations étant spécifiées à priori (il n'y a pas de requête en entrée du système). Par exemple, pour la question : "démocratisation CAUSE égalité", les réponses souhaitées sont sous forme d'extraits de segments textuels comme :"la démocratisation des sociétés entraîne une égalité des chances.", "L'égalité est favorisée par tout processus de démocratisation.", "Une démocratisation favorise une meilleure égalité.", etc.

Malgré leurs différences, les techniques de recherche d'informations et d'extraction d'informations se révèlent complémentaires. Leur association possède en effet un fort

potentiel dans la création ou l'amélioration d'applications d'extraction de connaissances à partir de textes. Il existe plusieurs moyens de combiner ces deux systèmes (Even, 2005) :

- Utiliser la recherche d'information en prétraitement de l'extraction d'information : face à un très large volume de textes, elle peut fournir à un système d'extraction d'information une sous-collection ne regroupant que les documents les plus pertinents.
- Des techniques propres à l'extraction d'information peuvent également être employées afin de compléter les approches classiques de recherche d'information pour catégoriser, filtrer et ordonner les documents en fonction de leur pertinence.
- Utiliser l'extraction d'information pour affiner les résultats d'un système de recherche d'information en améliorant la phase de modélisation des documents.

Dans le cadre de notre travail, ces deux techniques peuvent se suivre dans un modèle plus vaste : Des techniques d'extraction d'information sont employées à la suite de l'application d'une approche classique de recherche d'information et ce afin de catégoriser, filtrer et ordonner ou encore traiter les connaissances à fournir à l'utilisateur.
Pour ceci, nous considérons les hypothèses suivantes :

- Un texte contient des informations linguistiques explicites qui guident la lecture et qui suggèrent des "points de vue" de fouille sémantique,
- Ces informations peuvent être identifiées dans un texte ; elles conduisent à des annotations automatiques,
- Ces annotations automatiques sont exploitables pour extraire des informations dans une recherche d'informations, les catégoriser automatiquement, et pour les exploiter par des moyens opérationnels.

2.5 Enjeux et problématiques de la recherche d'objets pédagogiques à partir de documents textuels

Le World Wide Web initialement conçu pour une utilisation humaine contient des informations lisibles par des machines, mais pas forcément compréhensibles par elles. Il est difficile d'automatiser les traitements sur le web. Pourtant le volume d'informations contenues sur la toile est tel que cette automatisation devient nécessaire et le manque de structuration des données se fait ressentir.

De nouveaux standards comme XML ou RDF tentent de corriger e manque de structuration. XML (eXtensible Mark-up Language) est un langage mis en place par le W3C (World Wide Web Consortium) qui permet la description de documents électroniques par l'intermédiaire de DTD (Document Type Définition). Son but est de faciliter la diffusion d'informations sur l'Internet, l'accès dynamique aux bases de données et l'échange normalisé entre sites. C'est un langage étiqueté qui permet de structurer des données à l'intérieur d'un fichier texte. Par exemple, il permet de stocker des tableaux, des fichiers clients, des annotations, etc. il existe d'autres langages de données structurées mais ils sont souvent difficiles à utiliser. XML possède l'avantage d'être un standard et qui plus est destiné à l'Internet.

Des milliers d'apprenants et d'enseignants intègrent l'utilisation du Web et d'Internet dans la formation en vue d'apprendre et d'améliorer leur pratique pédagogique. Les usages effectifs correspondent davantage à une vision de l'informatique où l'ordinateur devient un moyen d'accès à l'information et à des ressources pédagogiques variées, un lieu d'échanges et de communication au sein d'une communauté d'apprenants (Pernin, 2003) où cohabitent technologies numériques mais également supports et outils traditionnels. Face à la croissance exponentielle des ressources pédagogiques (cours, exercices, études de cas, résumés,...) sur le Web, il convient de recourir à des procédés de description des données permettant l'extraction des objets pédagogiques à partir des documents.
L'indexation est l'un des nombreux procédés de description des données dont la finalité est de permettre une recherche efficace des informations contenues dans une collection de documents sans avoir à analyser chaque texte de document à chaque interrogation ou recherche (Goker et al., 2009). L'indexation est une technique qui permet à la fois de ranger d'une certaine manière mais surtout de retrouver rapidement les informations.
Nous donnons comme exemple le travail de (Dinh et al., 2010) qui propose une indexation des dossiers médicaux de patients. Nous présentons dans ce qui suit l'annotation faisant partie d'un processus préalable à l'indexation. Ensuite, nous illustrons quelques systèmes d'indexation des documents pédagogiques basée sur l'annotation.

2.5.1 Annotation préalable à l'indexation des documents pédagogiques

L'annotation est une des principales techniques utilisée pour indexer les documents textuels électroniques et plus particulièrement les documents pédagogiques (Gillard, 2002). Le terme "annotation" réfère à une note, une critique, une explication ou encore à un commentaire. Et puisqu'on ne critique qu'une idée ou un sujet, une annotation est alors obligatoirement liée à l'un de ces éléments ou à l'autre(Amardeilh, 2007).

Puisque nous nous situons dans le cadre du web sémantique, nous nous intéressons particulièrement à l'annotation sémantique que nous détaillerons dans la section suivante.

2.5.1.1 Annotation sémantique

Dans le cadre du web sémantique, un cas particulier existe pour l'annotation : c'est l'annotation sémantique. Il s'agit d'ajouter des notes aux ressources web pour qu'ils soient compréhensibles par les machines et par les êtres humains, ce qui permet de rendre exploitable les données échangées sur le web, tant par les êtres humains que par les machines. Cela permet également une meilleure interopérabilité entre les machines elles-mêmes, entre les êtres humains eux-mêmes et entre machines et êtres humains. Ces annotations ont été déjà utilisées dans plusieurs applications, comme La classification, la recherche d'information, le résumé automatique, l'interopérabilité.
L'annotation est dite sémantique si elle propose une solution qui répond aux besoins des utilisateurs dans le cadre du web. Elle peut être (Amerdeilh, 2007) :

- manuelle : elle consiste alors simplement à mettre en place une interface utilisateur dans laquelle l'utilisateur humain peut sélectionner la ressource à annoter, choisir le modèle formel servant à la création des annotations sémantiques et, tout en respectant les contraintes imposées par le modèle formel, créer les annotations voulues sur la ressource sélectionnée.

- semi-automatique : apporte une aide non négligeable à l'annotateur humain. Ces traitements semi-automatiques s'appuient généralement sur un moteur d'extraction d'information.

- automatisé : ce type d'annotation est en fait plutôt semi-automatique. En effet, les traitements entièrement automatisés utilisent des algorithmes basés sur des modèles statistiques (Dingli, 2003), ou sur l'exploitation de la redondance dans un corpus de ressources.

L'annotation sémantique permet de nombreuses applications comme la recherche d'information sémantique (Denoue et al., 1999), la catégorisation, la composition de documents, l'échange de documents (Daoust et al., 2006), etc. L'annotation sémantique est applicable à n'importe quel type de contenu : pages web, documents textuels non structurés, champs d'une base de données, documents audio ou vidéo, etc. Enfin, plus le modèle de l'annotation est formalisé, plus les services proposés à partir de cette annotation peuvent devenir "intelligents" (Amerdeilh, 2007).
Les annotations sémantiques sont définies par plusieurs caractéristiques (Desmontils et al., 2002) :

- Elles sont persistantes tant que le document initial n'est pas modifié.
- Elles sont implicites puisqu'elles s'ajoutent dans le code source du document mais attaché au document initial.
- Elles font généralement référence à une connaissance déjà existante dans une ou plusieurs ontologies.
- Elles sont opérationnelles du fait qu'elles sont censées être traitées par des machines. A contrario, des annotations en langage naturel qui ne sont compréhensibles que par les êtres humains.
- Elles ne sont pas consultables directement sur le document. Il faut posséder un éditeur qui vous permet d'accéder au code source du document et d'accéder donc aux annotations.
- Elles permettent de désambigüiser des termes polysémiques

L'annotation sémantique est alors une solution qui répond aux besoins des utilisateurs dans le cadre du web en termes (Prié et al., 2004) :

- D'optimisation de la recherche d'information sur les documents web.
- D'obtention d'informations sur les documents existants sur le web.

En résumé, la tâche d'annotation pour le web sémantique consiste à prendre en entrée des textes et à fournir en sortie le même texte enrichi par des annotations sémantiques basées sur la représentation de la connaissance plus ou moins formelle. Son objectif est de décrire le contenu des ressources en les annotant avec des informations non ambiguës afin de favoriser leur exploitation. Nous citons quelques travaux d'annotation, comme l'outil d'annotation SYDoM (Roussey et al., 2002) qui est un outil d'annotation pour le web sémantique. (Michelson et al., 2004), (Michelson et al., 2007) proposent aussi une méthode d'annotation des relations temporelles dans des sources non-grammaticales et non-structurées.
L'annotation des objets pédagogiques est un champ d'application pour plusieurs outils comme par exemple dans les travaux de (Lowe et al., 1997), les auteurs proposent une approche basée sur les "frames" pour l'annotation sémantique des objets pédagogiques. De par le passé, les objets pédagogiques ont été annotés par de multiples systèmes pour différentes finalités (Christiansen et al., 2004). Ces systèmes ont tous une spécificité en commun : alimenter les objets pédagogiques par des annotations en vue de les indexer plus tard ; et ce pour extraire les objets pédagogiques.

Reprenons une des particularités des annotations sémantiques, celle qui sont créées à partir de connaissances disponibles dans une ou plusieurs ontologies.

Une ontologie est une spécification formelle, explicite et consensuelle de la conceptualisation d'un domaine. Elle représente à la fois cet objet de consensus pour les êtres humains et un objet formel permettant son exploitation par un agent logiciel. Elle se compose de classes, d'attributs et de relations qui définissent et précisent l'utilisation de ces classes et de ces attributs. Elle est décrite dans un langage formel de représentation des connaissances (Exemple : RDF, OWL,...) (Amardeilh, 2007).

RDF (the Resource Description Framework) est un formalisme de représentation des connaissances, issu des réseaux sémantiques, dont la syntaxe utilise XML. Il sert à décrire des ressources documentaires par un ensemble de métadonnées (auteur, date, source, descripteurs, etc.).

OWL (Ontology Web Language) permet de formaliser une ontologie, ou plus globalement des ressources terminologiques et ontologiques, par la définition des concepts utilisés pour représenter un domaine de connaissance. Ce langage permet de décrire ces concepts par un ensemble de propriétés, de relations et de contraintes. Le formalisme utilisé correspond aux propriétés de certaines logiques de description(Kiryakov et al., 2004).

Les ontologies fournissent les moyens d'exprimer les concepts d'un domaine (Amardeilh, 2007) :

- Il s'agit de définir des concepts et de les relier par des relations sémantiques en premier lieu, de réaliser des modèles conceptuels en deuxième lieu, et aussi de dessiner des graphes conceptuels.
- Une ontologie se constitue principalement de concepts (classes, objets abstraits, objets concrets, objets réels,...), de relations entre ces concepts et d'attributs relatifs à ces concepts.
- Les instances de concepts font partie de la base de connaissances.
- L'action de définir et d'instancier une base de connaissances est appelée "peuplement d'ontologie", alors que "l'enrichissement d'ontologie" consiste à ajouter des concepts, des attributs ou des relations à l'ontologie existante.

Nous soulignons le fait que les deux termes "Annotation" et "Métadonnées" ont le même sens pour certains, cependent pour nous la différence se situe au niveau de l'élément auquel est affectée l'annotation : les métadonnées sont relatives au document, en entier comme l'auteur, la date de création d'une ressource, le titre, etc. Alors que "les annotations" viennent décrire le contenu, généralement segmenté, d'un document.

Plusieurs travaux se sont intéressés aux ontologies, comme (Boucetta et al, 2008), qui proposent un appariement sémantique des documents à base d'ontologie pour le e-recrutement. (Desmoulins et al., 2000) utilisent les ontologies pour indexer des documents techniques pour la formation professionnelle. Des ontologies sont aussi utilisées par (Dgim et al., 2006)
Dans le domaine pédagogique, les enseignants et les apprenants sont souvent contraint à traiter de grands volumes de données provenant de diverses sources documentaires. A partir de l'ensemble des documents qu'ils sont chargés d'étudier, ceux-ci doivent d'abord sélectionner les ressources documentaires pertinentes pour leur travail puis, pour chacune d'entre elles, extraire manuellement l'information pertinente. Cette information peut être ensuite annotée par un ensemble d'annotations et peut servir donc à une indexation des documents.
Dans ce type d'indexation, il s'agit d'utiliser la structure sémantique qui peut être construite par les annotations pour identifier des éléments dans un document et naviguer des éléments sémantiques aux fragments de texte ou vice-versa. Les usages de la représentation sémantique issue de l'annotation peuvent, à grands traits être séparés en deux classes. Dans la première, l'annotation est utilisée pour sa valeur sémantique, et sa source textuelle n'est plus nécessaire une fois que le travail d'analyse a été fait. Il s'agit alors d'extraire des connaissances, éventuellement d'alimenter des bases de données (Par exemple, une annotation effectuée pour le peuplement d'une ontologie de domaine ou pour remplir un formulaire concernant des informations sur une personne). Dans la seconde, l'annotation sert à accéder au texte qui lui a donné naissance. Nous parlons alors d'indexation sémantique (Nous citons l'exemple de notre système qui est un système de recherche et d'extraction des informations pédagogiques à partir de documents répondant à une requête utilisateur).
Cependant nous soulignons, l'utilisation d'une ontologie linguistique, dans notre travail, et non pas d'une ontologie de domaine pour annoter et indexer les objets pédagogiques dans les documents textuels. Dans la section suivante, nous présentons quelques systèmes d'annotation d'informations pédagogiques en vue de leur indexation.

2.5.1.2 Systèmes d'annotation des informations pédagogiques en vue de leur indexation

Il existe des systèmes d'annotation conçus et implémentés en vue d'indexer les documents et en extraire les informations pédagogiques. Nous présentons dans ce qui suit quelques uns de ces systèmes :

- **Le système QBLS**

Il s'inscrit dans le cadre des logiciels informatiques d'apprentissage. Il s'agit précisément d'un système d'apprentissage d'un cours à des étudiants par un enseignant ou par les étudiants eux-mêmes. Le contenu d'un cours est divisé suivant les types de savoirs-clés : Cours (sujet du cours), Question (exercices de travaux dirigés relatifs à ce cours), Notion (Concepts du domaine), thème (problématique clé). A Chacun de ces savoirs-clés est associé un ensemble de fiches contenant (Dehors et al., 2005) :

- L'Enoncé, Procédure de résolution de la question et Solution relatifs au savoir clé "Question"
- Définition, Exemple, Précision, Formalisation relatifs au savoir clé "Notion"
- Problématiques clés relatifs au savoir-clé "Thème"

Chacune des fiches relatives aux savoirs-clés "Question" et "Thème" peut contenir des "Notions" dont chacune a des liens avec les fiches Définition, Exemple, Précision ou Formalisme de cette notion (cf. Fig.2.3).

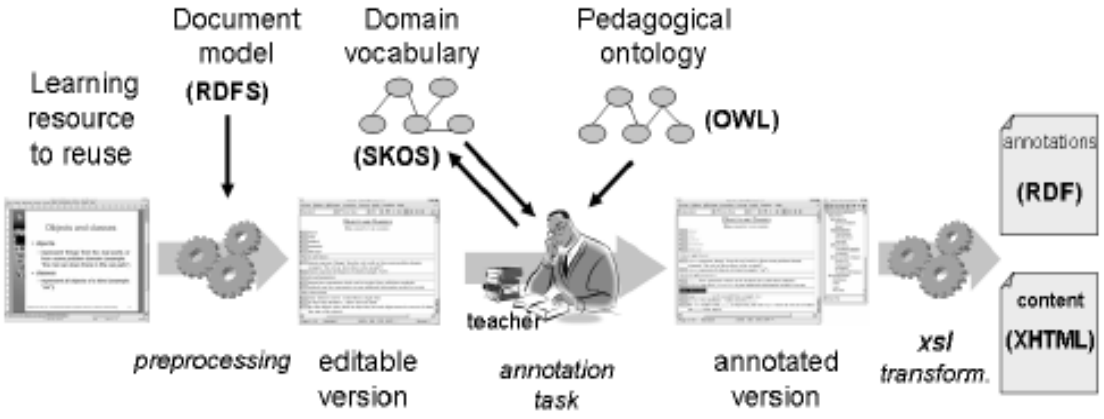


FIGURE 2.3: Méthode d'annotation de QBLS (Dehors et al., 2005)

Pour arriver à structurer un cours initial dans un modèle structuré, il doit être annoté manuellement par l'utilisateur (enseignant) relativement à la mise en forme de chaque paragraphe (ex : exemple ou définition). L'annotation est réalisée à l'aide du langage RDF en se référant à une ontologie pédagogique contenant les éléments d'un cours. Elle est structurée suivant la description faite ci-dessus. Les ressources déduites à partir du cours (document) initial sont stockées avec leurs annotations respectives dans un "entrepôt de connaissances pédagogiques". Une fois que le cours a été annoté sémantiquement, le système offre à l'utilisateur (enseignant ou étudiant) plusieurs services dans le cadre de l'apprentissage d'un cours :

- Pour l'étudiant : Le système propose un ensemble de questions à l'étudiant dans le cadre du cours et lui fournit la procédure de résolution de ces questions et la solution. Ainsi qu'il lui offre un détail (Définition, Exemple, Précision ou Formalisation) de la notion à détailler pour une meilleure compréhension.
- Pour l'enseignant : Le système permet à l'enseignant de présenter son cours de manière à faciliter la navigation entre les différentes parties de ce dernier (thème, notion, questions,...). Un cours présenté en parallèle avec des travaux dirigés est également possible, facilitant ainsi l'acquisition des connaissances pour l'étudiant.

Quand l'utilisateur pose sa requête indiquant par exemple une notion ou un thème, le moteur de recherche Corese s'active pour rechercher et afficher les fiches relatives à la requête posée. Seules les fiches "prioritaires" seront affichées. Une règle est indiquée pour le moteur de recherche pour qu'il distingue la fiche prioritaire de la non-prioritaire : si la fiche est une fiche de "Définition", ou d' "énoncé" ou d'une " problématique-clé" alors la fiche est prioritaire sinon elle ne l'est pas. Dans le cas où le moteur ne trouve pas de fiche prioritaire, il affiche la première fiche non prioritaire. Plusieurs expérimentations ont été appliquées sur le système QBLS concluant que le système nécessite encore plus d'améliorations pour satisfaire ses utilisateurs de profils différents.

- **Le projet TRIAL SOLUTION**

  C'est une plateforme d'aide à la production de support d'enseignement ou d'apprentissage suivant le profil utilisateur (enseignant ou étudiant). Son rôle est d'annoter des livres électroniques en vue d'en extraire des ressources pédagogiques, et de permettre par la suite aux enseignants et aux étudiants de faire des recherches sur ces ressources annotés sémantiquement (Buffa et al., 2005). Ces ressources sont stockées sur un serveur accessible pour tous les utilisateurs (cf. Fig.2.4.

L'outil d'annotation des ressources pédagogiques inclus dans ce projet permet de corriger et d'améliorer les annotations introduites automatiquement à ces ressources au cours de la première étape du projet. La correction et l'amélioration doivent être accomplies par un expert ou un connaisseur du domaine pour garantir une bonne amélioration. Ces annotations englobent le contenu sémantique d'une ressource, leurs interrelations dans le livre d'origine et l'auteur de celui-ci. Ce qui fait que les améliorations peuvent comporter : l'attribution d'un titre aux ressources pédagogiques et l'édition de leurs contenus, ainsi que l'édition de la structure arborescente de l'ensemble de l'entrepôt des ressources. Ces annotations se

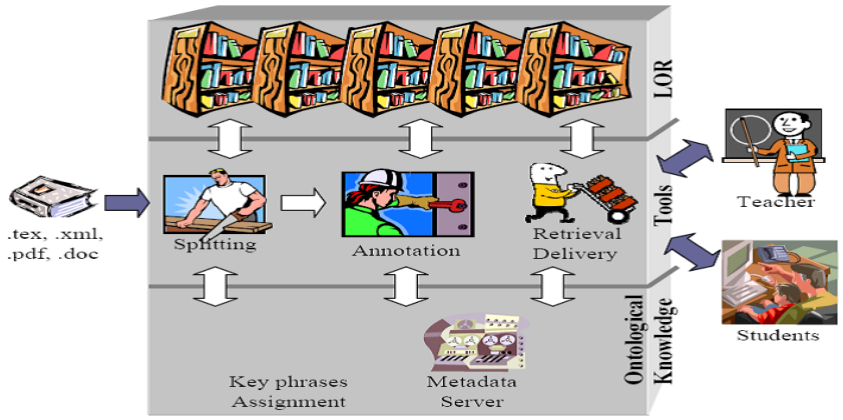


FIGURE 2.4: La plateforme TRIAL SOLUTION (Buffa et al., 2005)

basent sur un thesaurus d'un domaine (celui des mathématique) qui lui aussi peut être modifié par ce même logiciel d'annotation. Une modification concerne l'ajout de termes au thesaurus, description de ces termes, ajout de relations entre ces termes, etc.Les annotations relatives aux ressources pédagogiques sont conformes aux standards Dublin Core Metadata, IMS Learning Resousrce Metadata, LOM Learning Object Metadata.

- **Le système SYFAX**

  C'est un système d'informations de nature pédagogique s'intéressant au domaine de l'informatique. Il est situé sur le web et accessible par toute communauté d'utilisateurs (universitaire ou autre) (Smei et al., 2005). SYFAX propose une recherche des documents pédagogiques, une gestion des documents, un filtrage collaboratif des documents et une communication entre les différents membres de la communauté

  Le filtrage des documents commence suite à une extraction des documents à partir du web et à une annotation manuelle de ces documents par les différents utilisateurs. Chaque document pédagogique est annoté selon (1) sa correspondance avec le profil utilisateur (Oui/Non), (2) son type (TD, TP, etc.) à partir de l'ontologie "Type de documents" qui a été créée manuellement, (3) le point de vue de l'utilisateur sur le document (intéressant, moyen, peu intéressant), et (4) les concepts du domaine traités par le document en se référant à une ontologie du domaine de l'informatique construite automatiquement à partir d'un dictionnaire informatique nommé FOLDOC. L'annotation des documents peut être manuelle (effectuée par l'utilisateur), semi-automatique (prise en compte de l'avis de l'utilisateur) ou automatique (effectuée par un système d'extraction automatique des

métadonnées introduites par l'utilisateur). Le filtrage se fait selon la correspondance des documents au profil de la communauté. Le résultat est une base de données contenant des documents et leurs annotations respectives.
La recherche d'informations s'effectue à partir d'une base de ressources pédagogiques grâce à un moteur de recherche basé sur les annotations introduites, sur le profil utilisateur ainsi que sur l'usage d'ontologies. La requête de l'utilisateur subit un processus de raffinement pour distinguer les types et les sujets des documents recherchés. Il s'agit d'une expansion basée sur l'utilisation d'une ontologie des types des documents pédagogiques et une autre des domaines des documents informatiques. Les documents pertinents sont ceux de même type que la requête ainsi que ceux traitant les mêmes concepts que ceux énoncés dans la requête utilisateur.

Les systèmes proposés ont abordé le problème de recherche d'informations selon divers angles : soit par la structuration manuelle du cours selon une ontologie pédagogique dans le but de l'exploiter dans un environnement e-learning, soit par l'annotation manuelle des documents en vue de la production de supports de cours. Dans tous les cas, une intervention humaine est requise afin d'enrichir les documents par des annotations. Cependant, plusieurs producteurs de ressources pédagogiques ne s'intéressent probablement pas au retour aux documents pour annoter leurs propres travaux.

- **Le travail de (Hassen et al., 2009)**

  Il consiste à comparer l'efficacité des algorithmes Naïve Bayes et SVM (Salton Vector Machine) pour la classification des ressources pédagogiques en se basant sur un ensemble de propriétés (catégorie du contenu, titre du cours, année, auteur, etc.). C'est à l'utilisateur d'annoter le matériel pédagogique par des valeurs qui instancient des propriétés relatives à des objets d'apprentissage. Ces propriétés sont : Niveau éducatif (chaque page doit être annotée selon sa valeur éducative), Pertinence de la page (Quatre niveaux de "non-pertinent" à "très pertinent"), la catégorie du contenu d'un objet pédagogique (Définition, Exemple, Illustration, etc.), le type de la ressource (page web, encyclopédie, blog, forum, livre en ligne, etc.) et l'expertise de l'annotateur dans chacun des sujets sélectionnés selon quatre niveaux.

- **Le travail de (Thompson et al., 2003)**

  Il s'intéresse à la recherche de ressources pédagogiques à partir du web en procédant à une classification des ressources selon leurs types (Travaux Dirigés, Programme, Travaux Pratiques), ensuite à une extraction d'informations relatives

à ces ressources (Titre du cours, Auteur, Date, etc.). Les ressources représentent des pages HTML ayant quatre types : Travaux Dirigés, programmes, tutoriels et examens. Ces ressources sont liées à des cours du domaine de l'intelligence artificielle et l'apprentissage automatique, et sont téléchargés à partir du web grâce à un moteur de recherche. Les auteurs accèdent à ces ressources, à partir de la page initiale du cours, grâce à des liens hypertextes. Pour la tâche de classification, la pertinence des résultats est calculée selon une classification "Two-class" et une autre "Multi-class" selon les quatre types cités ci-dessus. Pour la tâche d'extraction, les auteurs utilisent un outil d'annotation pour annoter les programmes avec cinq métadonnées : Numéro du cours, Titre du cours, Auteur, Année. Ensuite, le modèle de Markov caché est appliqué pour extraire un label de la classe représentant le nom du champ à extraire. Toutefois, le but de ce travail est limité à une extraction de métadonnées relatives au document en entier en vue de les classer.

- **Le travail de (Meyer et al., 2007)**

  Ils proposent une classification des articles extraits de Wikipédia selon leurs catégories (thèmes) en utilisant la méthode du K-plus proches voisins. L'utilisation de cette méthode est argumentée par l'hypothèse l'hiérarchisation des catégories (catégories et sous catégories), ainsi que la dépendance entre les différentes catégories.

D'autres approches basées sur des patrons linguistiques ont été appliqués dans plusieurs travaux pour annoter et extraire les définitions à partir de ressources pédagogiques, par exemple :

- **Le travail de (Muresan et al., 2002)**

  Ils ont développé une grammaire capable d'identifier les patrons linguistiques définitoires de différents types. Pour ce faire, ils ont détecté les principaux patrons présents dans le corpus et ont écrit les règles appropriées pour leur extraction. La méthode appliquée est concrétisée dans le système DEFINDER.

- **Le travail de (Westerhout et al., 2008)**

  Ils ont appliqué les patrons linguistiques afin de constituer un glossaire. Les auteurs ont développé une grammaire capable d'identifier les patrons linguistiques relatifs aux définitions. Les auteurs ont annoté manuellement 21 fichiers appartenant à des contextes définitoires pour détecter des définitions de plusieurs types (être, verbe, pronom, etc.). Cependant, ces patrons ne peuvent pas détecter le type "mise en page" qui se base principalement sur des ponctuations (: / ; / .).

- **Le travail de (Greenwood et al., 2004)** Pour répondre à divers types de questions, ils ont développé des patrons linguistiques :

  - Questions à un argument nécessitant comme réponse un autre argument (ex : Tom Cruise est marié à qui ?)
  - Questions à liste nécessitant plusieurs arguments pour répondre à la question (ex : Listez 16 compagnies qui produisent des voitures)
  - Questions de définition où la réponse est sous forme textuelle couvrant une description du terme à définir (ex : “C’est quoi l’aspirine”).

  A chaque type de question est associé un ensemble de patrons linguistiques. Ensuite, à chaque patron linguistique relatif à une question est associé un patron de réponse à cette question. A travers les travaux présentés, nous remarquons que les patrons linguistiques sont appliqués la plupart du temps pour extraire des objets pédagogiques de type ”Définition“ en raison de l’accessibilité des structures langagières relatives à ce type que ce soit sur le web (Wikipédia, dictionnaires, etc.) ou dans d’autres sources comme les rapports, les manuels d’utilisation, etc.

- **Le travail de (Liu et al., 2003)**

  Ils proposent une méthode permettant à l’utilisateur d’apprendre davantage sur un sujet à partir du web. Les techniques proposées commencent par identifier les sous-thèmes ou les concepts similaires au thème entré par l’utilisateur. Ensuite, il informe les pages informatives, contenant des définitions et des descriptions du thème et des sous-thèmes comme c’est le cas dans un livre classique. La technique proposée procède par les étapes suivantes :
  Soit T une phrase soumise au système représentant le thème recherché

  - Soumettre T dans un moteur de recherche qui retourne un ensemble de pages pertinentes
  - Le système recherche les sous-thèmes et les concepts similaires à T à partir d’un ensemble S des pages classées en premier lieu par le moteur de recherche, et ce en utilisant des techniques du Data Mining.
  - Le système découvre les pages contenant les définitions du thème T, des sous-thèmes de T et les concepts similaires à T. Les définitions sont repérées en appliquant les patrons linguistiques utilisées dans le système DEFINDER présenté dans une section précédente.

Dans ce qui suit, nous ajoutons aux systèmes présentés précédemment deux systèmes (Karina et ALOCoM) qui présentent des points en communs avec notre travail.

- **Le système Karina**

  Le travail de (Chabert-Ranwez, 2000) a participé à la réalisation de deux projets fortement similaires, Karina et Sybil. Le principal objectif de ces projets, c'est la composition de documents structurés en sélectionnant, organisant et assemblant des fragments de documents électroniques (briques d'information). Les deux approches consistent à exécuter des requêtes dans une base de données, en fonction d'objectifs pédagogiques précis, puis à composer un document cohérent à l'aide des informations extraites de la base. Le document ainsi obtenu est un document multimédia adapté à un utilisateur donné. Les deux méthodologies suivent les étapes de base de la composition : recherche d'informations, filtrage, ordonnancement et assemblage. Nous détaillons dans ce qui suit le projet Karina qui représente le principal travail de la thèse en question.

  Les principales phases du projet Karina sont détaillées dans ce qui suit :

  - Attribution des rôles pédagogiques aux briques d'informations

    Un rôle pédagogique est un attribut associé à une brique d'information dont la valeur traduit sa capacité à susciter un certain comportement chez l'apprenant. Au cours du projet, plusieurs rôles sont présentés ainsi que l'interaction correspondante suscitée chez l'apprenant (cf. Tab. 2.1).

    Les rôles que jouent des briques d'informations au sein d'une narration sont décisifs dans le processus de composition. La recherche d'informations et le filtrage en sont directement dépendants car pour homogénéiser un discours, il faut disposer d'un nombre adéquat de définitions, d'explications, d'exemples, etc. L'ordonnancement est, quant à lui, directement dépendant des rôles. Les briques d'informations traitées dans ce projet sont des documents HTML homogènes, annotés à l'aide d'un outil de qualification, selon une ontologie du domaine. L'attribution des rôles est effectuée en appliquant la théorie des sous-ensembles flous vu que ce domaine contient des approximations, des incertitudes, des imprécisions, des nuances. L'inconvénient majeur de cette approche réside dans le fait que les rôles pédagogiques associés aux briques d'information soient statiques au sein de la qualification.

  - La recherche d'informations

    Suite à une requête de l'apprenant, le système extrait les briques d'information susceptible d'être incluses dans le document final et dont la qualification est à une distance sémantique acceptable de la requête de l'apprenant.
    Une brique d'information qualifiée est considérée d'un point de vue éditorial,

Rôle pédagogique	**Action correspondante attendue de l'apprenant**
Appariement	Mise en correspondance de certaines entités
Conclusion	Lecture
Description	Lecture et assimilation
Définition	Lecture et assimilation
Exemple	Lecture et assimilation
Exercice cas général	Résolution exacte
Exercice d'ordonnancement	Ordre exact
Exercice de prononciation	Prononciation acceptable
Explication	Lecture et assimilation
Formule	Lecture et assimilation
Illustration	Lire, écouter, regarder en fonction du média
Introduction	Lecture, compréhension du contexte
Questions à choix Multiple	Choix exacts
Référence	Lecture de l'ouvrage référencé
Résolution de problème	Résolution exacte
Résumé	Lecture
Test vrai/faux	Sélectionner les propositions
Théorème	Lecture et assimilation

Tableau 2.1: Rôles pédagogiques et actions attendues de l'apprenant (Chabert-Ranwez , 2000)

où les auteurs, la date de création, les autorisations d'utilisation de cette brique, le format, la langue, etc. sont spécifiés. Ensuite les métadonnées portant sur la totalité du document sont décrites (élément global). Un élément segment permet d'annoter des parties de briques ayant une portée définie (par sa dimension textuelle s'il s'agit d'un texte ou sa dimension temporelle s'il s'agit d'une vidéo ou d'un enregistrement). Enfin, il est possible de qualifier des évènements ponctuels par des descriptions locales.

Pour formuler ses objectifs, un apprenant sélectionne à partir d'une interface un objectif pédagogique correspondant au but de son étude. Pour chacun d'eux, le système recherche dans la base de données les briques d'information dont la qualification est à distance sémantique acceptable, en utilisant des requêtes SQL. Pour chaque brique d'information, le système recherche les briques d'informations correspondant à ses pré-requis.

- Le filtrage

  Dans cette étape, seule la dimension temporelle est prise en compte. Les briques d'information d'une façon générale possèdent une information temporelle indiquée dans le champ durée de leur description : le temps de lecture pour un texte, la durée pour un enregistrement audio ou vidéo, le temps de prise de connaissance pour une image. Dans Karina cette durée correspond au temps de lecture de l'hypertexte définit d'une manière approximative car il

est fortement dépendant du lecteur, du contexte de lecture et des interactions de l'utilisateur. Le système fait la somme des durées des briques d'information sélectionnées et la compare à la période de temps disponible de l'utilisateur.

- La méthode d'organisation

  A ce stade de la composition, le système dispose d'une liste de briques d'informations à organiser. Pour ce faire, il utilise les contraintes imposées par les pré-requis entre les briques d'information exprimées dans leur qualification. Si plusieurs objectifs sont mentionnés dans la requête sans lien explicite entre eux, le système traite chaque objectif séparément et assemble les documents obtenus dans l'ordre d'énonciation des objectifs.

- L'assemblage et la présentation

  Les briques d'information sont présentées à la suite les unes des autres, sans transition entre elles. Par contre, un assemblage physique est réalisé, puisque les briques d'informations sont accessibles au travers de l'interface.
  L'approche adoptée dans Karina et la nôtre présentent plusieurs points communs comme l'utilisation du langage XML pour qualifier des documents, la description sémantique du contenu de ces documents, la description globale d'un document mais également sa description par fragments ou bien sa description locale, etc.

- **Le système ALOCoM**

  Le système ALOCoM (Verbert et *al.*, 2005) commence par segmenter le contenu des diapositives de la présentation en titres, paragraphe, listes, images, diagrammes et tables. Dans une deuxième étape, les différents composants sont catégorisés en utilisant les patrons textuels : définitions, exemples, références, introductions et résumé. Une troisième étape consiste à annoter les différents composants par des métadonnées en utilisant la plateforme "Automatic Metadata Generation". Cette étape facilite le repérage des composants pertinents.
  Le système ALOCoM est intégré dans l'application MS PowerPoint. Ainsi un enseignant voulant créer une nouvelle présentation peut rechercher des définitions, diapositives, exemples, références et images qu'il veut réutiliser.
  Les composants répondent à la requête de l'enseignant sont affichés et peuvent être ajoutés aux diapositives de la présentation. Le système ALOCoM applique une approche pour extraire les définitions basée sur l'approche appliquée dans DEFINDER, présenté précédemment. Il utilise une technique basée sur les règles pour extraire les définitions des articles médicaux. Il utilise des phrases comme : "is the term for", "is defined as", "is called" ainsi que des marqueurs de textes comme par exemple : (–, ( ))

2.5.1.3 Synthèse au sujet de l'indexation des objets pédagogiques basée sur l'annotation

Pour résumer les travaux présentés, nous illustrons un tableau (cf. Tab.2.2) qui réunit plusieurs paramètres, essentiels à notre égard, à l'annotation des informations pédagogiques contenus dans les textes. Ces paramètres sont :

- Type de l'annotation (manuelle, semi-automatique, automatique),
- Type du document annoté (Livres, cours, articles wikipédia, etc.),
- Contenu des annotations,
- Méthode adoptée pour l'annotation,
- Utilisation ou non d'une ontologie de domaine,
- Finalité de l'annotation.

TABLEAU 2.2: Synthèse des différents travaux sur l'annotation des documents pédagogiques

Paramètres / Travaux	**Type d'annotation**	**Type docs à annoter**	**Contenu des annotations**	**Méthode appliquée**	**Utilisation ou pas d'une ontologie**	**Finalité de l'annotation**
QBLS	Annotation manuelle et semi-automatique	Support de cours (Word et Latex)	- Sujet du cours - Exercices relatifs au cours - Concepts du domaine - Problématique clé	- Sélection semi-automatique des métadonnées à partir d'une ontologie	Utilisation d'une ontologie du domaine pédagogique	Offrir différentes stratégies de navigation dans les ressources pédagogiques d'un cours donné
TRIAL SOLUTION	Annotation manuelle et semi-automatique	Livres électroniques (Word et Latex)	- Type de la ressource (Définition, exemple, etc.) - Mots clés - Relations avec les autres ressources	Sélection semi-automatique des métadonnées à partir d'une ontologie	Utilisation d'une ontologie du domaine des mathématiques	Construire un entrepôt de ressources pédagogiques
SYFAX	Annotation manuelle ou semi-automatique	Principalement HTML et XML (Document structurés)	- Correspondance du document avec le profil utilisateur (Oui/Non) - Son type (TD, TP, etc.) à partir de l'ontologie - Point de vue de l'utilisateur sur le document (intéressant, moyen, très intéressant) - les concepts du domaine traités par le document	- Sélection manuelle des métadonnées. Le système peut intervenir pour compléter des données en cas où il manque des annotations	Utilisation d'une ontologie du domaine pédagogique et une autre du domaine des mathématiques	-Recherche des documents pédagogiques -Gestion des documents -Filtrage collaboratif des documents -Communication entre les différents membres de la communauté

Hassen et Mihalcea	Annotation automatique de la propriété éducative d'un document	Page web Livres en ligne Blogs Présentation Publication	-Valeur éducative -Pertinence - Catégories du contenu (Définition, Exemple, illustration, Question, etc.) - Type de la ressource (Page web, livres en ligne, tec.) - Expertise de l'annotateur	Une saisie semi-automatique des métadonnées et application des algorithmes Naive Bayes et SVM pour la classification		Classification des ressources éducatives selon leurs valeurs éducatives
Thompson et al	Annotation semi-automatique	Pages HTML	Numéro du cours, Titre du cours, Auteur, Année	Sélection semi-automatique des métadonnées		Classification des ressources selon leurs types pour une recherche éventuelle
Meyer et al	Annotation automatique	Articles WIkipédia	-Thème et sous thèmes de l'article	K-plus proches voisins		Classification des articles selon leurs thèmes
DEEFINDER	Annotation automatique des segments	HTML, PDF, Doc (Convertis en XML)	-Définition de type : être, verbe, ponctuation, mise en page, pronom	Les patrons linguistiques		Recherche des définitions
Westerhout et al	Annotation automatique des segments	HTML, PDF, Doc (Convertis en XML)	-Définition de type : être, verbe, ponctuation, mise en page, pronom	Les patrons linguistiques		Construction d'un glossaire
GreenWood et Saggion	Annotation automatique des	Questions et Réponses	Type de la question (Question à une seule réponse,	Les patrons linguistiques		Répondre à des questions

	segments		Question à liste, Question de définition)			
Liu et al	Annotation automatique des segments	Pages informatives	Thèmes et sous-thèmes	Les patrons linguistiques		Apprendre davantage sur un sujet à partir du web
KARINA	Annotation du rôle pédagogique d'une brique d'information	Documents HTML homogènes	Conclusion, Description, Définition, Exemple, Exercice, Explication, Formule, Illustration, etc.	Annotation en utilisant un outil de qualification basé sur une ontologie du domaine	Utilisation du domaine pédagogique	Recherche et filtrage de briques d'informations, Assemblage et Présentation des briques d'information
ALOCoM	Annotation des composants d'un fichier	Fichier PowerPoint	Définitions, diapositives, exemples, références et images	Les patrons textuels		Rechercher des composants textuels et aide à constituer sa présentation PowerPoint

D'après cette synthèse, nous pouvons souligner les limites de ces systèmes : quantité d'information à stocker et à retrouver, cohérence dans le cas où un grand nombre d'utilisateurs annoterait les pages, problèmes de confidentialité, non-intégration de ces systèmes dans les navigateurs commerciaux.
De plus, la plupart des travaux de recherche portant sur l'indexation de documents pédagogiques se sont intéressés à une indexation du document en l'annotant par un ensemble de métadonnées relatives à la totalité du document, alors qu'une analyse détaillée du contenu textuel du document s'avère nécessaire pour répondre aux requêtes utilisateur.
Pour d'autres systèmes, une intervention humaine est requise afin d'enrichir les documents par des annotations. Cependant, plusieurs producteurs de ressources pédagogiques ne s'intéressent probablement pas au retour aux documents pour annoter leurs propres travaux.
D'autres travaux sont contraints d'être appliqués sur un domaine bien particulier, en se référant à une ontologie de domaine, comme le domaine des mathématiques, qui nécessite des ressources énormes pour sa construction.
A notre connaissance, plusieurs des travaux présentés précédemment se sont intéressés à une indexation des documents en les classifiant selon un ensemble de propriétés. Si ces travaux semblent être capables d'indexer les documents en les classant avec efficacité, elles sont cependant limitées, notamment pour une extraction des connaissances, ainsi qu'une indexation sémantique des textes.

D'autres approches plus linguistiques proposent l'utilisation de nouvelles méthodes d'identification automatique de relations sémantiques discursives associées à des segments textuels variés. Elles font appel à une analyse linguistique et à une meilleure compréhension de l'organisation discursive des textes. L'intégration de cette nouvelle méthode d'annotation, dans un système de recherche d'informations pédagogiques, reste prometteuse. La démarche que nous adoptons s'inscrit dans cette lignée et sera détaillée plus tard dans cette thèse. En plus, diverses expériences ont montré que l'amélioration de la performance des systèmes d'indexation passe par l'intégration d'au moins deux modèles - linguistique et statistique.
Nous présentons dans ce qui suit divers systèmes d'indexation des documents pédagogiques basés sur des techniques statistiques.

2.6 Conclusion

Nous avons débuté le chapitre par une discussion des concepts d'extraction d'information et extraction de connaissances. Ensuite nous avons présenté l'annotation comme technique d'indexation des documents pédagogiques. Nous nous sommes intéressés particulièrement à cette technique car elle sera appliquée plus tard dans notre modèle de recherche d'objets pédagogiques. Des systèmes appliquant l'annotation comme technique d'indexation, sont par la suite illustrés comme solution au problème de l'extraction des informations pédagogiques. La majorité de ces systèmes, bien qu'ils répondent à une bonne partie des besoins des utilisateurs, présentent quelques problèmes critiques tels que l'obligation de l'intervention humaine dans le processus d'annotation, ou encore le recours à des ontologies de domaine dans ce même processus d'annotation. En plus, la plupart de ces systèmes se limitent à l'extraction des *définitions* à partir des documents. Dans le chapitre suivant, nous présentons notre modèle de recherche d'objets pédagogiques à partir de documents, qui est un modèle basée sur l'annotation sémantique de ces objets.

Chapitre 3

Modèle d'extraction d'objets pédagogiques à partir de documents

Sommaire

3.1 Introduction

Face à la croissance exponentielle des documents pédagogiques (supports de cours, exercices, études de cas, etc.) sur le Web, il convient de recourir à des procédés d'extraction et de recherche d'informations utilisés par le plus grand nombre. Si la recherche d'informations à partir de documents textuels à l'aide de mots-clefs offre des performances intéressantes en termes de rapidité de traitement, ses résultats ne sont pas directement exploitables et nécessitent un travail considérable d'analyse des documents sélectionnés pour extraire l'information pertinente.
Les approches d'extraction automatique appliquées jusque-là sont basées essentiellement sur des comptages statistiques de co-occurrences de mots-clefs (Stapley et al, 2000), (Pillet, 2000) ou sur des règles ou automates d'extraction définis manuellement, à base de verbes significatifs et des termes linguistiques (Blaschke, 1999), (Thomas et al, 2000), (Poibeau, 2002). Les résultats obtenus présentent, soit une précision très faible, soit une couverture limitée. L'extraction automatique d'informations pédagogiques à partir de documents nécessite donc la mise en œuvre de méthodes d'extraction d'informations plus complexes qui s'appuient sur la sémantique. L'aspect novateur de notre approche réside dans la conception et l'implémentation de techniques informatiques originales basés sur des connaissances linguistiques présentes dans les textes et structurées sous forme de types d'objets pédagogiques, selon la méthode d'exploration contextuelle (Desclés, 2006).
Notre système est un système d'extraction d'objets pédagogiques, qui doit offrir un certain nombre de fonctionnalités à l'utilisateur. Il fonctionne selon trois modalités successives mais indépendantes : la première consiste à annoter d'une manière automatique et sémantique les objets pédagogiques situés dans les documents du corpus. La deuxième vise à assister l'utilisateur (apprenant ou enseignant) dans la recherche d'informations pédagogiques en lui offrant des réponses pertinentes sous la forme d'objets pédagogiques. La troisième prend en compte les réponses à la requête de l'utilisateur pour en constituer une fiche pédagogique rassemblant les objets pédagogiques pertinents.
Ce chapitre présente notre méthodologie pour l'annotation sémantique et automatique des objets pédagogiques, ensuite leur extraction à partir des documents. Une fois assemblés, ces objets forment une fiche pédagogique répondant à la requête utilisateur. Il est organisé comme suit : dans la section 1, nous commençons par rappeler brièvement le contexte d'utilisation de notre système. Nous donnons ensuite des détails sur notre corpus de travail dans la section 2. La section 3 sera consacrée à la présentation du principe de la méthode d'exploration contextuelle et ses applications. Dans la section 4, nous présentons notre modèle pour l'extraction d'objets pédagogiques à partir de documents textuels, en détaillant ses différentes phases.

3.2 Contexte d'utilisation

Le contexte d'utilisation a été abordé dans le premier chapitre, cependant, nous rappelons quelques éléments nécessaires à la compréhension de notre système. L'implication de l'utilisateur dans un système est liée aux "contextes d'utilisation du système". Le contexte représente l'environnement et les contraintes d'utilisation du système, à savoir :

3.2.1 Qui utilise le système ?

La connaissance de l'utilisateur et du cadre d'utilisation sont primordiaux afin de spécifier les compétences et les présupposés sur lesquels se fondera le système. Se demander qui utilisera le système revient à dresser un panorama des compétences des utilisateurs afin de borner les demandes qu'on peut leur adresser. L'effort de prise en main et de formation face à l'outil doit être limité.
Les utilisateurs de notre systèmes mis au point sont en général des apprenants et des enseignants, ou des personnes ayant des besoins pointus en recherche d'informations pédagogiques dans un domaine bien déterminé. Ces utilisateurs sont censés pouvoir juger de la validité d'une information en contexte. Ils ne sont pas obligatoirement des linguistes ou des informaticiens.
Nous mettons les trois types d'utilisateurs "étudiant", "apprenant" et "enseignant" dans le cadre de notre étude, pour cela nous optons pour les définirons du dictionnaire Larousse, qui sont comme suit :

- Etudiant : personne qui fait des études supérieures dans une université ou un établissement d'enseignement supérieur, une grande école.
- Apprenant : personne qui suit un enseignement quelconque.
- Enseignant : personne dont le métier est d'enseigner

3.2.2 Quelles sont les interactions entre l'utilisateur et le système et quels services le système doit-il fournir ?

L'interaction entre l'utilisateur et le système est un point essentiel dans le cadre d'une application de recherche d'informations. Les systèmes d'extraction d'informations offrent le plus souvent des possibilités d'interactions assez limitées, contrairement à notre système qui offre plusieurs services à l'utilisateur. Nous précisions que notre système n'est pas destiné à servir que des utilisateurs dans les domaines de l'enseignement et de

l'apprentissage, ce qui fait que les interactions entre l'utilisateur et le système dépassent largement le cadre pédagogique.
Cette étude vise la mise au point d'un système opérant sur des documents principalement pédagogiques. Notre système doit offrir un certain nombre de fonctionnalités selon le contexte d'utilisation. En effet il doit viser dans un premier temps à assister l'utilisateur dans la recherche d'informations pédagogiques en lui offrant d'autres fonctionnalités. L'ensemble des services offerts par notre système peut être résumé dans les points suivants :

- Segmentation du corpus choisi par l'utilisateur
- Annotation du corpus choisi par l'utilisateur
- Réponse aux requêtes formulées par l'utilisateur
- Constitution de fiches pédagogiques selon le besoin de l'utilisateur
- Gestion des règles d'Exploration Contextuelle

3.3 Corpus de travail

La communauté linguistique considère, à la suite de Sinclair (1996), qu'un corpus est "une collection de données langagières qui sont sélectionnées et organisées selon des critères linguistiques explicites pour servir d'échantillon de langage". D'après cette définition, un ensemble de données collectées ici et là sans réflexion préalable sur ce qui motive le rassemblement des documents n'est pas un corpus. Pour les tenants de la linguistique de corpus, il n'est pas possible de rassembler de façon exhaustive toutes les formes répondant à l'objet d'étude (clôture du corpus). D'après cette définition, toute collection de données, même expérimentales, peut être un corpus à part entière.

La définition, en anglais, proposée par (Gibbon et al, 1998) est la suivante : "A corpus is any collection of speech recordings which is accessible in computer readable form and which comes with annotation and documentation sufficient to allow the re-use of the data in_house, or by people in others organizations".

D'après (Habert et al., 1997), un corpus est une collection de données langagières qui sont sélectionnées et organisées selon des critères linguistiques et extralinguistiques pour servir d'échantillon d'emplois déterminés d'une langue. Les corpus sont généralement utilisés dans les buts de produire des applications commerciales (amélioration d'une application en cours de développement, tester la robustesse d'un outil) ou de les utiliser dans les recherches

L'utilisation d'un corpus nous amène à se poser les questions suivantes :

- Quel type de corpus ?
- Comment constituer le corpus ?
- Dans quel but l'utiliser ?
- Quelles annotations ?

3.3.1 Corpus constitué de documents pédagogiques

Le repérage d'objets pédagogiques pertinentes nécessite de cibler précisément les informations pédagogiques dans le cadre de documents principalement pédagogiques.
A certains types d'objets pédagogiques sont associées certaines sources d'informations privilégiées avec chacune leurs caractéristiques structurelles propres. Exemples :

- Le type d'objet pédagogique "Exercice", se focalise sur les travaux dirigés, les travaux pratiques, les examens, etc.
- Le type d'objet "Cours" est généralement repéré dans les supports de cours.
- Le type d'objet "Définition" s'intéresse aux différentes formulations possibles d'une définition dans un cadre pédagogique.

Concernant la méthode de constitution de notre corpus, il n'y a pas de méthode systématique. En fait, nous sommes allés chercher du Web et nous avons téléchargé des documents sous formats numériques. Nous avons également profité de documents pédagogiques chez nos collègues (supports de cours, examens, etc.), utilisés dans leur processus d'enseignement.

3.3.2 Description et caractéristiques de notre corpus de travail

Le corpus constitué présente un style assez pédagogique mais ne suit pas un format particulier ; les documents sont généralement identifiés par un auteur et un titre. Puisqu'il n'y a pas de structure particulière dans un document pédagogique, chaque document a sa structure particulière à lui. Mais généralement les documents pédagogiques contiennent un titre, des sous titres, leur contenu est décomposé de petits paragraphes. Ils contiennent des définitions, des exercices, des exemples, des comparaisons, etc. Nous avons choisi de travailler sur des documents qui contiennent ce type de données ou d'autres objets pédagogiques. Un grand nombre de ces documents sont disponibles sur le web, mais ils ne sont pas rédigés dans le but d'y faire de l'extraction d'informations.

Ce sont généralement des documents plus ou moins longs (3 à 30 pages), contenant des marqueurs qui indiquent les types d'objets pédagogiques recherchés. Actuellement, notre corpus est exclusivement constitué de documents en français. Nous présentons ci-dessous un exemple d'extrait de notre corpus :

Exercice 5
Rappelez brièvement :

1. *ce qu'est un pont et quelle est son utilité dans les réseaux Ethernet ;*
2. *quelle sont les responsabilités d'un pont (vis à vis d'un répéteur) ;*
3. *quel est le principe du pontage transparent et quel problème il pose.*

Pour notre travail de thèse, nous avons constitué deux corpus :Un corpus d'acquisition des données linguistiques et un autre d'évaluation.

3.3.2.1 Un corpus d'acquisition des données linguistiques

Nous avons constitué un corpus composé principalement de documents pédagogiques téléchargés à partir d'internet. Ce corpus couvre plusieurs domaines comme Base de données, Linguistiques, Littérature, Gestion, etc. et ce dans le but de montrer que notre modèle est applicable dans plusieurs domaines. La composition détaillée de notre corpus d'acquisition des données linguistiques (30 fichiers) est la suivante :

- 5 fichiers de type Microsoft Word.
- 20 fichiers de type PDF.
- 5 fichiers de HTML.

Ces fichiers ont une longueur moyenne de 10 pages.

3.3.2.2 Un corpus d'évaluation

Notre corpus d'évaluation est composé de 300 documents, principalement de nature pédagogique : Support de cours, travaux dirigés, travaux pratiques, présentations PowerPoint, Syllabus, et des documents de différentes natures. Ces documents sont des fichiers de différents formats (DOC, PDF, PPT, HTML, TXT, etc.) et ont une longueur moyenne de 10 pages.

3.4 Présentation de la méthode d'exploration contextuelle

Nous avons choisi de présenter cette méthode dans son cadre théorique, avant de décrire sa mise en œuvre dans le cadre de notre système. En effet, c'est une méthode qui a été appliquée dans plusieurs applications et qui a prouvé son efficacité tout au long de son cycle de vie.

3.4.1 Principes généraux

La méthode d'exploration contextuelle (EC)(Desclés et al., 1991 ; Desclés et al., 2007 ; Desclés et al., 2009) est une méthode générale de traitement du langage. Elle se propose d'apporter des solutions à des problèmes très divers liés au langage indépendamment d'un domaine particulier. C'est un ensemble de stratégies décisionnelles qui permettent de construire des représentations à l'aide d'un examen des éléments linguistiques à l'intérieur de leur contexte textuel.

Cette méthode considère qu'un texte n'est pas seulement l'expression d'un ensemble structuré d'objets et de concepts (prédicats unaires) décrits par des réseaux avec héritage, comme le laisseraient entendre certains travaux en recherche d'informations et en TAL, qui "effacent" avant tout traitement et exploitation des textes, toutes les "unités grammaticales" pour ne retenir que les "unités lexicales" qui seraient les seules unités porteuses de signification. Un texte est surtout le résultat d'opérations et de relations prédicatives grammaticales, ainsi que d'opérations discursives (de mise en texte) qui laissent des traces linguistiques clairement identifiables.

En général, un texte n'est pas une simple énonciation de faits, il exprime également des engagements ou désengagements de l'énonciateur, des jugements, des prises de position. Certes, certaines indications discursives peuvent être absentes dans un genre et spécifiques d'un autre genre mais les instructions de lecture et de structuration textuelle restent semblables. À côté des catégories grammaticales et lexicales d'une langue, les catégories discursives contribuent à une forte structuration sémantique des textes avec des mises en texte : organisations discursives, organisation des contenus, organisations dialogiques. Comme nous l'avons dit plus haut, les opérations de mise en texte laissent des traces dans les textes, il convient donc de les identifier pour reconstruire ces opérations et de chercher à les exploiter pour répondre à certains besoins du traitement automatique des textes, en particulier en annotant automatiquement certains segments textuels. L'annotation automatique à partir de ces traces conduit à un enrichissement endogène du texte, sans faire appel à des connaissances externes (par exemple à des

ontologies des domaines). Chaque texte contient donc en lui-même des indications sur sa propre structure et sur son contenu. Le texte annoté automatiquement permet de procéder à des applications avec des ressources linguistiques relativement économiques.

3.4.2 Notion de point de vue

La notion de point de vue est liée à celle de l'Exploration Contextuelle. Elle a pour but de découper l'espace de recherche en différents sous-espaces qui correspondent chacun à un type d'informations donné : informations quant à la cause, informations quant à la définition,etc. Notons que la notion de point de vue discursif est indépendante d'un domaine particulier ou d'une thématique particulière. Chaque point de vue est ainsi susceptible de s'appliquer à n'importe quel type de textes, quel que soit leur objet ou leur contenu. Il peut y avoir des points de vue qui dépendent du domaine. Il est bien évident cependant que les informations de nature pédagogique, par exemple, sont moins abondantes dans des articles de presse que dans des documents pédagogiques : mais ce type d'informations est néanmoins susceptible d'apparaître dans tous les types de documents.

Ainsi, l'idée qui motive la notion de point de vue est de rendre l'information du Web, ou plus généralement, des grandes banques d'informations, plus facile à identifier, en se focalisant sur un besoin spécifique, sans que l'utilisateur soit un expert du domaine traité (Laublet et al., 2002). De ce fait, le "rôle du point de vue" correspond au besoin à satisfaire par le point de vue choisi : recherche d'une explication causale, recherche des acteurs d'un domaine, recherche d'une définition, recherche d'un exercice, recherche d'un exemple et plus ambitieusement, arriver à satisfaire des besoins tels que : s'informer, apprendre, ou comprendre....

Dans le cadre de notre travail, la notion de point de vue est traduite par le type d'un objet pédagogique. Ainsi, quand un utilisateur donné cherchera à satisfaire son besoin par le choix du point de vue "Exercice", ce besoin est traduit par la recherche des objets pédagogiques de type "Exercice". Cette partie sera détaillée plus tard dans le chapitre.

3.4.3 Fouille sémantique par exploration contextuelle

La fouille sémantique sur les textes ne peut être entreprise automatiquement que si l'on s'appuie exclusivement sur des marqueurs linguistiques qui sont autant d'indications explicitement liées à des recherches comme par exemple : "identifier une hypothèse" ;

"identifier une définition" ; "de quoi parle un texte ?" ; "Qui parle à qui ?" ; "Qui cite qui ? Comment ?" ; "Qui a rencontré qui ? Où ? Quand ?".

Chacun de ces points de vue de fouille se voit associer des classes de marqueurs linguistiques généraux, indépendants des domaines et donc non spécifiques à un texte particulier. Ces marqueurs sont en fait des expressions (parfois discontinues) qui signalent une certaine information discursive, ils sont donc les traces linguistiques de certains points de vue de filtrage que l'on peut exploiter pour entreprendre une fouille sémantique des textes. À chaque classe de marqueurs est associé un ensemble de règles d'exploration contextuelle. En effet, un marqueur seul n'est souvent pas un indicateur suffisant pour identifier l'annotation sémantique qu'il conviendrait d'assigner au segment textuel dans lequel il apparaît. Il convient donc, pour réduire le bruit, de procéder à une recherche, autour d'un indicateur identifié, d'indices linguistiques complémentaires qui viendront soit confirmer la valeur discursive de l'indicateur identifié, soit infirmer cette dernière.

Les règles d'exploration du contexte doivent donc être formulées pour prendre une décision, en fonction de l'indicateur identifié qui déclenche la règle et des indices linguistiques figurant dans le contexte. À partir des conditions de la règle qui doivent être présentes, on annote le segment textuel où se trouve l'indicateur linguistique, selon le point de vue discursif dont il est la marque expressive. Ainsi, il est assez clair que le mot "défini" est un bon indicateur pour détecter dans un document pédagogique une définition. Cependant, pour attribuer au segment textuel l'annotation d'une définition, il est nécessaire de procéder à une exploration du contexte de "défini", pour vérifier cette annotation, information qui pourra être importante dans la recherhce d'informations ou la constitution automatique de fiches pédagogiques. Par exploration contextuelle, on tentera de vérifier la présence d'indices complémentaires dans le contexte immédiat comme : est-sont / par / ... De tels indices permettent d'affecter l'annotation à certains segments textuels.

Par exploration contextuelle autour d'indicateurs déclencheurs identifiés dans les textes, nous pouvons diminuer le bruit qui serait trop important si l'on opérait uniquement avec les déclencheurs, considérés comme des "mots clés" généralisés et comme des expressions régulières gérées par les automates finis. Un système d'exploration contextuelle (Desclés, 1993 ; Desclés, 2006 ; Desclés et al., 2009) est un ensemble de classes d'indicateurs linguistiques relatifs à un point de vue de fouille, accompagné d'un ensemble de règles d'exploration contextuelle, chaque règle étant spécifiée comme une règle de décision d'annotation :

SI une occurrence d'un Indicateur dans un segment linguistique (phrase, proposition...)

ET SI des occurrences d'indices linguistiques sont identifiées dans un espace linguistique de recherche défini autour de l'indicateur
ALORS annoter le segment textuel qui est déterminé par l'occurrence de l'indicateur

Nous présentons ci-dessous un schéma du processus de raisonnement dans une règle d'exploration contextuelle (cf.Fig.3.1) :

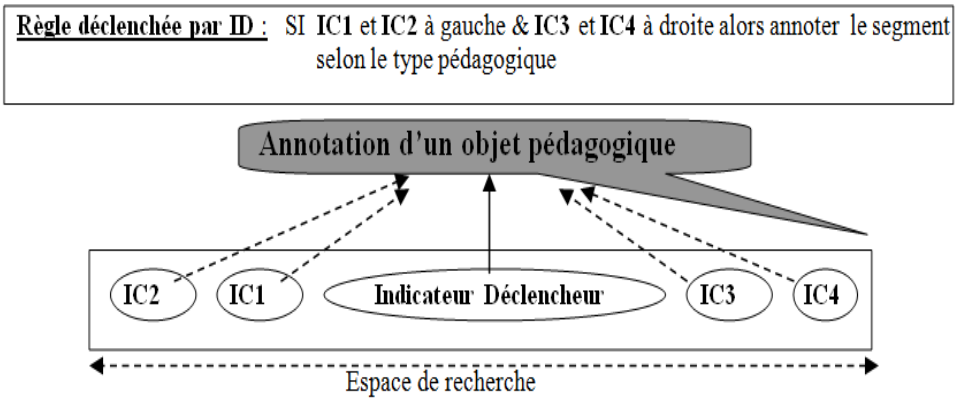


FIGURE 3.1: Schéma d'une règle d'exploration contextuelle (Desclés et Djioua, 2007)

Cette démarche nécessite un travail linguistique d'investigation dans les textes pour dégager les marqueurs discursifs, ainsi qu'une expertisation des ressources. Cependant, la constitution de ressources (marqueurs discursifs et règles d'exploration contextuelle) par des linguistes sur un point de vue de fouille précis ne nécessite pas dans le temps de lourdes ou continuelles modifications, mais seulement des réajustements. C'est la nature discursive des marques, comme nous l'avons dit, qui fait qu'elles restent stables et indépendantes des domaines et de leur évolution.

3.4.4 Le travail du linguiste

En amont de l'Exploration Contextuelle, le travail consiste d'abord à déterminer une carte sémantique des points de vue sur lesquels est menée l'étude. Construire cette carte sémantique revient à recenser et classer les différentes valeurs sémantiques qui expriment un point de vue donné et à les organiser en un réseau.
Une fois les valeurs sémantiques identifiées, la suite du travail consiste à collecter des marqueurs sémantiques et à observer la façon dont ils s'agencent dans le tissu discursif des textes : nous observons ainsi les morphèmes, les mots, expressions et locutions,

la typographie, enfin tous les moyens langagiers qui expriment potentiellement les valeurs sémantiques du type d'objet pédagogique. Ces marqueurs, dépositaires d'une ou plusieurs valeurs sémantiques, ont été considérés par (Desclés, 1997) comme des "indicateurs" (ou indices déclencheurs). Afin d'organiser et d'être en mesure d'automatiser ces processus dans des réalisations applicatives, ces différents principes ont été traduits dans un langage formel, qui a recours à la notion de "règles". L'Exploration Contextuelle est ainsi décrite par un ensemble de règles, qui sont déclenchées par l'identification d'un indicateur. Les règles diffèrent les unes des autres par les valeurs de leurs paramètres. Parmi ces différents paramètres, nous trouvons ainsi l'espace de recherche, appelé aussi contexte linguistique, qui est défini en fonction de la nature de l'indicateur. Les indices complémentaires seront également recherchés dans cet espace.

3.4.5 Applications de la méthode d'exploration contextuelle

Plusieurs travaux, au sein du laboratoire LaLIC, ont appliqué la méthode d'Exploration Contextuelle dans leur travaux pour identifier et interpréter les relations entre concepts (Le Priol, 2007), segmenter les textes (Mourad, 2001), annoter les entités nommées (Bouhafs, 2005), annoter les spécifications informatiques de besoins (Garcia-Flores, 2007), annoter les évènements (Elkhlifi et al., 2010), etc. Les principes, eux, ne diffèrent pas : une règle contextuelle est une description d'une configuration textuelle dans laquelle une liste d'indicateurs est représentative d'une des significations déclarées dans la carte sémantique. Cette configuration contextuelle comporte aussi l'ensemble des conditions sur les indicateurs et indices complémentaires. L'application de ces principes a donné lieu à plusieurs réalisations informatiques. Les résultats de ces réalisations ont grandement contribué à l'élargissement de la problématique de l'EC, en analysant à la fois la polysémie grammaticale, lexicale et l'annotation qui est passé des problèmes très précis de polysémie grammaticale à l'annotation automatique des notions sémantiques et discursives complexes (causalité, définition, résumé automatique).Plus récemment, impulsée par une nouvelle implémentation informatique de la méthode, la problématique s'est élargie à nouveau vers l'indexation sémantique des pages et la recherche dans le Web (Djioua et al., 2006).

D'après la thèse de (Garcia-Flores, 2007), nous distinguons quatre générations de systèmes d'EC :

- La première réalisation informatique de l'EC est le système SECAT (Desclés et al., 1991), destinée à la désambiguïsation des valeurs aspecto-temporelles dans le français. Ensuite le système expert SEEK (Desclés et al., 1993) a mis en place

la première architecture d'EC proprement dite, dédiée au repérage des relations statiques dans les textes et basée sur un moteur d'inférences et une base de connaissances. L'architecture de SEEK a inspiré les systèmes SAFIR et SERAPHIN (Berri, 1996a), tous deux consacrés au résumé automatique, ainsi que le système COATIS (Garcia, 1998) pour le repérage de la causalité.

– ContextO (Crispino et al., 1999) et Semantex (Ben Hazez, 2002) ont été conçus comme des plates-formes génériques d'EC et non comme des applications destinés à résoudre un problème spécifique. Le deuxième a été utilisé pour l'annotation des relations aspecto-temporelles (Chagnoux, 2003) et le premier pour la structuration des concepts (Le Priol, 2000), la causalité (Jackiewicz, 1998) et le résumé automatique (Minel, 2002).

– Le système d'EC est EXCOM 1 (Djioua et al., 2006). La rupture technologique avec ses prédécesseurs est importante et fonde la troisième génération d'outils d'EC. Parmi les travaux de cette génération, nous pouvons citer les travaux de (Bertin et al., 2006), (Blais et al., 2007), (Chai, 2007), (Le Priol et al., 2006). Nous citons aussi les travaux de (Elkhlifi et al., 2010) pour l'extraction des évènements.

– Le système d'EC le plus récent est EXCOM 2 (Alrahabi et al., 2010). Il se distingue par rapport à EXCOM 1 par une meilleure séparation entre indicateurs et indices. A chaque indicateur est associé un ensemble de règles d'Exploration Contextuelle, avec des indices positifs impliquant une annotation, et des indices négatifs annulant une annotation. Un processus d'exploration à partir de l'indicateur gauche ou droite, et non à partir des frontières externes du segment textuel à annoter potentiellement. Le moteur EXCOM 2 résulte d'une analyse systématique des "défauts" des systèmes précédents par Motasem AlRahabi et Jean-Pierre Desclés avec une réalisation informatique par Motasem Alrahabi dans sa thèse (Alrahabi, 2010). Le système EXCOM 2 a été mis en œuvre sur l'annotation automatique de textes en biologie (Recherche d'hypothèses plausibles) donnant lieu à des publications (Desclés, 2011), (Makkaoui, 2012). EXCOM 2 a été utilisé également pour l'analyse sémantique des références bibliographiques afin de répondre à des questions comme : "Qui a cité qui ?" et "Pourquoi ?" (Résultats, hypothèses, méthodes, etc.) (Atanassova, 2012) (Bertin, 2011).

Le tableau (cf.Tab.3.1) résume les caractéristiques des différentes générations.

Les principes de l'exploration contextuelle étant désormais décrits dans leurs grandes lignes, nous proposons sa description détaillée lors de son application dans notre système dans la section suivante.

Caractéristiques	**1ère génération**	**2ème génération**	**3ème génération**	**4ème génération**
Description des règles	Faite directement dans le langage de programmation du système	Faite dans un langage formel de description des règles	Faite en lanagage XML	Faite en lanagage XML
Segmentation des textes	Sans segmentation préalable des textes	Segmentation préalable des textes	Forme partie intégrante de l'architecture	Forme partie intégrante de l'architecture
Organisation des ressources linguistiques	Ne peuvent pas être réutilisées	Peuvent être réutilisées et organisées selon la notion de point de vue de fouille	Peuvent être réutilisées et organisées selon la notion de point de vue de fouille	Peuvent être réutilisées et organisées selon la notion de point de vue de fouille
Programmation	Programmation structurée	Programmation structurée	Programmation structurée	Programmation structurée
Architecture de l'application	N'est pas conçu pour un environnement Web	N'est pas conçu pour un environnement Web	N'est pas conçu pour un environnement Web	Orientée vers l'insertion dans un environnement Web pour une sélection de documents qui sont ensuite soumis à l'annotation
La langue de traitement des textes	Fançais	Fançais	Anglais Fançais	Anglais Fançais Arabe Koréen (partiel -lement)

Tableau 3.1: Illustration des générations des systèmes d'EC

3.5 Modèle proposé pour l'extraction d'objets pédagogiques

Entremêlant une perspective double, à la fois d'analyse linguistique et d'apprentissage automatique, notre modèle entend aborder la question de l'extraction des objets pédagogiques à partir des documents. Dans un premier temps, nous appliquons une analyse linguistique, dont l'objectif est de parvenir à une annotation des structures linguistiques relatives aux différents types d'objets pédagogiques. Ensuite, nous procédons à une indexation des documents basée sur les résultats de l'annotation.

Nous appliquons une méthode d'apprentissage automatique (représentation vectorielle), en plus de l'annotation, afin d'améliorer la pertinence des résultats par rapport à la requête utilisateur. Finalement, c'est l'index qui réunira les résultats de l'annotation et de la représentation vectorielle pour permettre par la suite l'extraction des objets

pédagogiques. En effet, la requête utilisateur est composée de deux parties : le type d'objets pédagogiques recherchés (Type d'objet pédagogique) et leur domaine (Terme à rechercher).

Le type de l'objet sera indexé en se basant sur les résultats d'annotation des objets par exploration contextuelle. C'est la première et principale composante de notre système puisque le type d'un objet pédagogique est décisif dans le processus d'indexation. L'extraction d'informations en est directement dépendante car pour homogénéiser un discours pédagogique, il faut disposer d'un nombre adéquat de définitions, d'exemples, d'exercices, etc. La deuxième composante qui est une composante secondaire est celle de la représentation vectorielle des objets utilisant la méthode de (Salton, 1970). La constitution de fiches pédagogiques est une composante optionnelle, mais très intérressante dans le système, vu qu'elle permet de réunir le contenu des objets sélectionnées dans un même document.

Toutes ces composantes constituent notre modèle (cf.Fig.3.2) qui sera détaillé tout au long de ce chapitre. Ce modèle est d'abord présenté dans la figure suivante. Ses principales composantes sont :

- L'annotation sémantique et automatique des objets pédagogiques, qui elle-même composée d'une sous composante de segmentation des documents et une autre d'annotation des objets.
- La représentation vectorielle des objets pédagogiques
- L'extraction des objets pédagogiques répondant la requête utilisateur
- La constitution de fiches pédagogiques

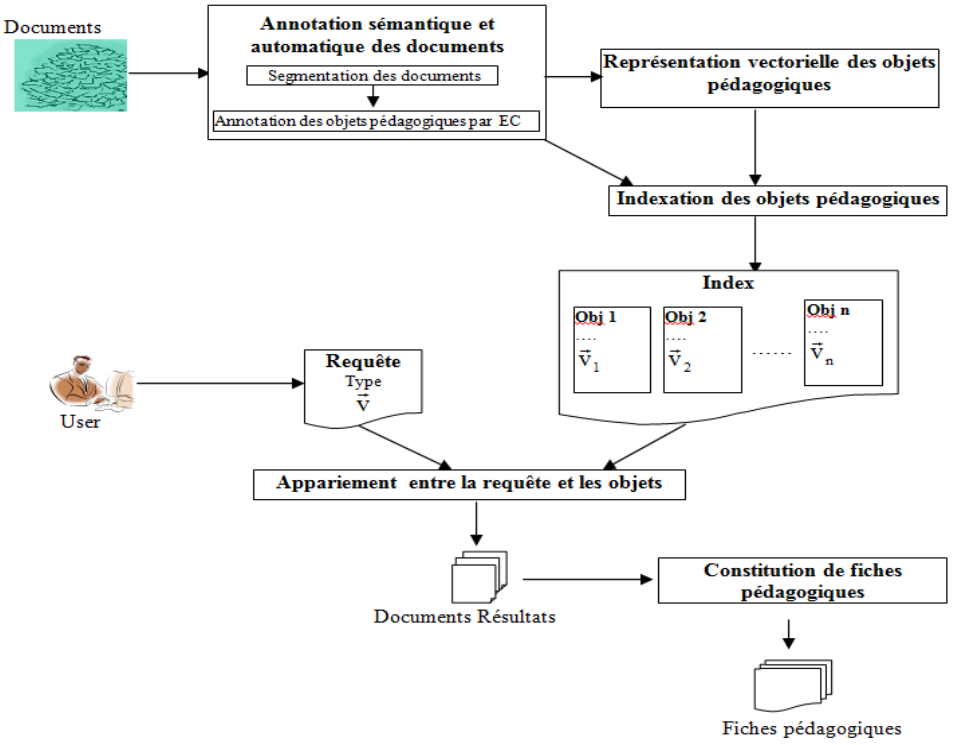


FIGURE 3.2: Modèle proposé pour l'extraction d'informations pédagogiques à partir de documents

3.5.1 Annotation sémantique et automatique des objets pédagogiques

La composante annotation est la principale composante dans notre modèle. Elle comporte des sous étapes de segmentation et d'annotation. Ces sous étapes appliquent principalement la méthode d'Exploration Contextuelle présentée au début du chapitre. Ces sous étapes seront détaillées tout au long de cette section.

3.5.1.1 La segmentation

La segmentation est définie par (Mourad, 2002)comme la détermination des limites des unités linguistiques dans un texte (unités comme proposition, phrase, paragraphe, etc.). Pour l'analyse sémantique des textes, il faut être capable de le segmenter en des unités linguistiques qui sont supérieures et inférieures à la phrase normative, en prenant en compte des marques sémiotiques clairement et formellement identifiables par une machine. Ainsi, la ponctuation et tous les indices typographiques restent les éléments les plus pertinents, car ils sont susceptibles de fournir des indications précises pour segmenter et structurer formellement les textes.

D'après (Bouhafs, 2005), nous distinguons trois types d'approches de la segmentation :

- Approches numériques (réseaux neuronaux, N-grammes, chaînes de Markov ...). Par exemple, dans l'outil SATZ (phrase en allemand), Palmer (Palmer et al., 1994) utilise un réseau neuronal, en étudiant par des critères lexicaux le contexte gauche et le contexte droit de chaque candidat (dans son cas les ".", "!", "?" définissent la fin d'une phrase).

- Approches par automates finis et expressions régulières : Pour la segmentation des textes français, Anne Dister a développé un segmenteur en utilisant le système INTEX (Silberztein, 1993). Pour procéder au découpage du texte en phrases, elle a appliqué un automate qui lit une séquence dans un texte et par rapport à l'information associée à celui-ci, l'automate insère la marque de fin ou de non-fin de cette séquence. Les marques utilisées pour cette phase de segmentation sont les ".", "!", "?". La règle générale est appliquée (après la levée de certains cas d'ambiguïté) dans les cas où l'on rencontre la séquence "." ou "?" ou "!" suivit d'une majuscule. Pour résumer, nous disons que la plupart des outils de segmentation existants sont limités à la simple utilisation des marques de ponctuation ".", "!", "?", avec une étude des quelques cas d'ambiguïté sur des corpus bien déterminés, et une utilisation de dictionnaires de sigles. Il faut noter que leur segmentation des textes est facilitée par le fait qu'une documentation technique est très structurée, ce qui est rarement le cas dans les documents pédagogiques. Ce point est détaillé dans plus loin dans cette section.

- Approches par exploration contextuelle autour des marqueurs de ponctuation (exploration contextuelle exemple : La segmentation de textes en segment textuel se fait à partir d'une étude systématique des marques de ponctuation. La segmentation est basée premièrement sur des marques de ponctuation (marqueurs typographiques) ".", ";", " :", "!", "?", "⁰" (qui sont considérées comme des marques pivot pour le déclenchement des règles de segmentation), et deuxièmement sur une étude des contextes gauches et droites de ces marqueurs, ce qui permet de lever les ambiguïtés. Cette segmentation s'appuie sur l'application de la méthode d'EC (Mourad, 2001) dans le cadre de développement de l'outil SegATex. Nous citons aussi l'outil SEEK-JAVA (Bouhafs, 2005) appliquant aussi la méthode d'EC pour segmenter les textes en phrases. Il prend en compte dans sa segmentation plusieurs considérations : le saut de ligne est considéré comme un marqueur de fin de phrase, un point d'interrogation ou trois points de suspension placés au milieu d'une phrase ne génère pas plusieurs segments, un point suivant une abréviation ou l'initiale d'un prénom n'est pas considéré comme une marque de fin de segment.

Ces considérations de segmentation améliorent les résultats puisque l'exploration doit avoir lieu dans le bon contexte.

Notre travail de segmentation s'inspire fortement des travaux effectués appliquant l'EC. Toutefois, pour les besoins de notre approche, nous avons reformulé et implémenté les règles de segmentation tirées de la thèse de G. Mourad. D'une manière générale ici le but est de poursuivre une segmentation à l'aide des balises <section>, <paragraphe> et <phrase> grâce à un ensemble de règles d'exploration contextuelle formulées en XML. Le résultat de la segmentation est un fichier XML ayant cette forme :

```
<article>
 <section>
  <title> Titre </title>
  <paragraphe>
   <phrase> Phrase 1 </phrase>
   <phrase> Phrase 2 </phrase>
  </paragraphe>
 </section>
</article>
```

3.5.1.2 L'annotation sémantique et automatique des objets pédagogiques

Dans ce module, nous procédons, d'abord, par une étude linguistique et théorique de chaque type d'objet pédagogique. Cette étude comporte, en premier lieu, une étude linguistique qui vise à circonscrire et éclairer le plus possible la spécificité et la nature de chaque type d'objet, sur laquelle il sera possible d'asseoir un repérage automatique des structures linguistiques relatives aux différents types d'objets. En deuxième lieu, une recherche des marqueurs saillants porteurs des types d'objets doit être effectuée : en parvenant à établir une première mouture de la carte sémantique des types d'un objet pédagogique, nous serons en mesure de commencer à ordonnancer les marqueurs principaux qui portent la trace de chaque type d'objet pédagogique. Ces deux premières sous-étapes ont en fait davantage été entreprises de concert, que l'une après l'autre, car la carte sémantique des types d'objets pédagogiques s'élabore progressivement, à mesure que nous parvenons à organiser les marqueurs qui portent la notion.

Suite à cette étude linguistique, une mise en œuvre opératoire de l'étude linguistique est effectuée pour une annotation sémantique des types d'objets dans les textes :

– Constitution de règles pour extraire de façon automatique des structures linguistiques relatives à chaque type d'objet pédagogique : Les marqueurs recueillis précédemment seront organisés dans des listes (regroupant soit des éléments de

même nature, soit des flexions de ces éléments, soit encore des groupes de synonymes), avec comme objectif de rendre ces regroupements les plus explicites et les plus opératoires possibles. Ces listes fourniront la matière première de règles d'"Exploration Contextuelle" (Desclés et al., 1993 ; Desclés, 2006), au moyen desquelles nous fouillons un texte pour en extraire des objets pédagogiques. Ces règles opèrent une hiérarchie entre des indicateurs saillants et des indices complémentaires : elles cherchent, dans le contexte d'un indicateur fort, des indices permettant d'établir la valeur sémantique dont il est porteur.

– Evaluation des règles : Ce premier ensemble de règles sera soumis à une évaluation qui cherchera notamment à mesurer, dans les énoncés annotés comme relevant de chaque type d'objet, le bruit (énoncés extraits non pertinents) et le silence (l'absence d'extraction là où une annotation manuelle en apporterait). De toute évidence, le bruit sera considéré comme beaucoup plus gênant que le silence, puisqu'il s'agit surtout d'un premier ensemble de règles destiné à être enrichi. Les règles doivent donc être réutilisables ultérieurement.

- **Etude linguistique des objets pédagogiques : Points de vues et types d'objets**

  L'une des motivations majeures de l'idée de points de vue, c'est la volonté de mettre en évidence et d'identifier d'une façon ciblée dans les textes les informations les plus importantes et les plus indicatives d'une manière différenciée. Force est de constater que l'exploration du web met l'utilisateur face à tous types et sources d'informations. Cette diversité des types d'informations traitant d'un sujet constitue pour les utilisateurs (enseignants et les apprenants) une richesse informationnelle sur un sujet donné. La question qui se pose alors est la suivante : Comment traiter ces informations avec toutes les sources d'informations sans se restreindre à un domaine d'investigation ou à un corpus particulier ?

  La notion de point de vue peut répondre à cette question ; puisque chaque point de vue cible dans les sources d'information une trace informationnelle particulière captant un aspect particulier de l'information recherchée. La notion de point de vue peut être alors comprise à partir des requêtes exprimées par l'utilisateur à propos d'un sujet particulier. Le choix d'un point de vue peut permettre de faciliter l'expression de ce besoin. En effet, il suggère une orientation pour la recherche ou une piste de réponse à ces besoins.

  Pour illustrer le type de recherche que nous cherchons à rendre possible, imaginons un utilisateur (enseignant ou apprenant) qui cherche sur le Web à s'informer ou à apprendre à propos de la notion de *maintenance*. Il peut par exemple commencer par chercher une définition, comme première étape, ensuite il va essayer,

comme deuxième étape, de chercher des exemples, des exercices sur cette notion. Les exemples suivants montrent des extraits pouvant être obtenus, suivant les marqueurs utilisés, pour les points de vue respectivement de la définition et de l'exemple :

- "*La maintenance est définie comme l'ensemble des activités destinés à maintenir ou à rétablir un bien dans un état de sûreté de fonctionnement.*"
- "*Le nettoyage d'une machine est un exemple de maintenance d'une machine.*"

La pluralité des points de vue présents dans notre méthode ouvre la voie à de nouveaux modes de recherche de l'information sur le web. L'utilisateur dont le besoin souvent complexe est difficile à expliciter, va pouvoir ainsi le cerner et l'exprimer à travers sa requête. Les points de vue de notre système de recherche d'objets pédagogiques sont traduits en termes de types d'objets pédagogiques. les différents besoins des utilisateurs ainsi que leurs types d'objets pédagogiques respectifs et leurs sous-types.

Afin de proposer à l'utilisateur un système d'extraction d'objets pédagogiques à partir de textes, nous avons étudié le contenu des documents pédagogiques, qui développent explicitement le type d'information recherchée, à savoir des supports de cours, des supports de travaux dirigés, des supports de travaux pratiques. Ainsi, nous avons fait le constat suivant : plusieurs informations liées à un objet pédagogique, ne dépendant pas d'un domaine spécifique, peuvent être fréquemment employées. Par exemple, le type d'objet, son sous-type, son contenu, son emplacement, etc. tentent de décrire un objet que l'on cherche à circonscrire. Par ailleurs, ces notions permettent aussi d'introduire des informations qui peuvent intéresser fortement l'utilisateur du système. Ainsi, elles peuvent être exploitées afin de constituer des fiches pédagogiques personnalisées selon le besoin de l'utilisateur.

Notre méthode s'appuie sur ces types d'objets qui introduisent des informations pour les utilisateurs apprenants. Chaque type d'objet s'exprime dans les textes pédagogiques par l'intermédiaire de plusieurs termes, par exemple, le type "Exercice" est exprimée au moyen des marqueurs tels que : *répondez, proposez, etc.*, ou par des marqueurs nominaux tels que : *exercice, Questions à choix multiples, Questions, etc.* Toutefois, pour identifier cette notion, la présence de certains indices complémentaires (*à, une, des, etc.*) dans un espace contextuel de recherche peut s'avérer nécessaire. En effet, pour chaque terme exprimant un type d'objet, plusieurs formes flexionnelles et plusieurs contextes doivent être pris en compte pour identifier l'information introduite. La reconnaissance de ces schémas est réalisée à l'aide de la méthode d'exploration contextuelle (Desclés, 1997), (Berri, 1996b),

(Minel, 2000) (Desclés et al., 2009). Notre hypothèse est que chaque type d'objet pédagogique laisse des traces discursives dans le document texte. Les types d'objets pédagogiques sont décrits comme suit :

- D'une part, une relation complexe entre les concepts dans une structure "carte sémantique" et d'autre part un ensemble de classes et sous-classes d'unités linguistiques (indicateurs et indices).
- Un ensemble de règles d'exploration contextuelle associées à une classe d'indicateurs ayant un fonctionnement équivalent.

- **Approche Contextuelle des différents types d'objets**

  La notion de "type" joué par un objet pédagogique est d'autant plus difficile à définir qu'un type peut dépendre de plusieurs paramètres : le contexte dans lequel l'objet est utilisé, la forme qu'il prend et bien sûr son contenu. Plusieurs découpages sont possibles selon le rôle ou l'usage potentiel de l'objet (fragment). Dans le système (Chabert-Ranwez, 2000), le critère est le rôle pédagogique qui a pour but de susciter l'action pédagogique chez l'apprenant. Ainsi, les fragments (appelées briques d'information) sont caractérisées selon leurs rôles (une conclusion, une description, etc.).
  Dans l'approche proposée par (Nestorov et al., 1997), le terme employé est "rôle" pour désigner un point de vue particulier sur un objet (Définition, Conclusion, Résumé, Illustration, etc.). Notre travail partage la notion de point de vue particulier sur un objet tout en lui donnant comme désignation le mot "type".
  De ce fait, par le mot "type", nous désignons le rôle pédagogique d'un objet envers un apprenant ou un enseignant, comme : *Définition, Exemple, Exercice, Plan, etc.*

  - Objet d'étude

    Pourquoi entreprendre une étude et un repérage des types d'objets pédagogiques (Définition, Exercice, Exemple, Plan, Caractéristique) ? Nous avons déjà mentionné que le recours aux exemples, aux exercices, par exemple, caractérise aussi bien nos discussions quotidiennes les plus triviales que les productions textuelles où une visée pédagogique ou scientifique est nettement perceptible. Leur rôle dans la structuration des documents pédagogiques, permettant à la fois son intelligibilité et sa maîtrise, les rend indispensables pour la production de supports de cours, de travaux dirigés, de travaux pratiques, etc.

    En outre, dans une perspective d'échange des connaissances, les objets pédagogiques occupent une place importante. Il est ainsi possible d'entrevoir des applications au repérage de ces objets pédagogiques, notamment dans l'optique

d'une amélioration des systèmes d'extraction d'informations. Il faut bien mesurer ici l'intérêt d'une étude linguistique des objets en vue de leur repérage automatique. En effet, l'accessibilité des objets pédagogiques devient une problématique de plus en plus vive, à laquelle des acteurs de plus en plus nombreux comme les enseignants, les étudiants et les apprenants sont confrontés. Face à ces besoins et ces demandes, certains chercheurs s'efforcent de fonder leurs recherches non plus seulement sur une approche des textes par mots-clés, dont on commence à percevoir les limites, mais sur une approche qui tienne plus finement compte des contenus mêmes des documents pédagogiques. Pour le cas précis qui nous retient, celui de requête portant sur des objets pédagogiques ayant différents types, le procédé est peu ou prou le même, bien que la spécificité de ces requêtes appellent un traitement un peu différent.

A rebours de l'approche de ce service, qui au fond consiste surtout à repérer des termes (mots-clés) répondant à la requête, on peut envisager de chercher les objets pédagogiques ailleurs, là où elles s'élaborent et où elles trouvent naissance : dans les textes eux-mêmes, dans les supports de cours, dans les livres, dans l'ensemble des productions textuelles. L'idée est plutôt d'aller chercher les objets pédagogiques là où elles prennent corps : dans les textes mêmes.

Ce qui est donc visé, c'est de ramener les objets pédagogiques depuis le lieu où ils se créent ; c'est d'extraire les objets pédagogiques développés et proposés par tel ou tel auteur, tel ou tel collectif, etc.
L'idée première est de s'atteler aux textes réels, et non pas construits pour illustrer une théorie. L'angle d'analyse qui nous retient, dans notre étude, est celui des traitements linguistiques mis en œuvre lorsqu'un objet pédagogique est identifié. L'approche consistera d'abord à parcourir le texte et repérer des connaissances représentatives des objets pédagogiques qui sont déposées en nombre (supports de cours, travaux dirigés, etc.) avant d'entreprendre une évaluation sur d'autres corpus. Les différentes valeurs sémantiques que revêt chaque type d'objet pédagogique seront ainsi illustrées par des énoncés tirés de textes réels et leurs principaux marqueurs seront explicités.

Cependant, si nous procédons à une analyse linguistique détaillée des objets pédagogiques, nous serons confrontés à de nombreuses questions. Quelles informations relatives aux objets pédagogiques retenir, pour savoir comment

différencier les données recueillies ? Quels marqueurs choisir pour pouvoir filtrer, interpréter et exploiter au mieux les données relatives aux différents types d'objets déposés dans les textes ? Les types d'objets Exemple, Exercice, Définition, méthode, Plan, etc. font appel à de nombreuses catégories grammaticales et se déploient à travers une large diversité de procédés linguistiques. Il est donc nécessaire de parvenir à s'en forger une conception claire, pour savoir quels indicateurs linguistiques retenir, quels autres écarter pour l'annotation automatique, notamment quand nous nous limitons au contexte pédagogique. Plusieurs conceptions relatives à chaque type d'objet ont été développées, mais ces types d'objets présentent une grande complexité théorique que cette étude n'entreprendra pas d'épuiser.

Pour chaque catégorie de la carte sémantique, nous avons défini l'ensemble des règles couvrant toutes les formes linguistiques possibles relatives aux objets pédagogiques. Nous commençons par un exemple textuel pour généraliser ensuite toutes les structures linguistiques. Cette méthode permet de définir d'une manière incrémentale une base de règle solide. En effet, nous donnons à l'utilisateur la possibilité de gérer la base de ses règles en ajoutant, supprimant ou modifiant des règles. L'étude linguistique des différents types d'objets pédagogiques donnera lieu ensuite à une approche directement opératoire, s'intéressant à l'une des retombées que l'on peut attendre d'un repérage automatique des objets pédagogiques : l'extraction automatique des objets pédagogiques.

- Polysémie des indicateurs des types d'objets pédagogiques

  Pour la plupart des types d'objets pédagogiques, les marqueurs linguistiques sont souvent polysémiques, même les plus usuels, comme le verbe 'désigner' pour le type "Définition'. Cette polysémie justifie ainsi pleinement l'approche contextuelle. L'exemple de la phrase :

  "Le maître d'école désigne un élève pour aller au tableau"
  "Le symbole X désigne un des constituants de la phrase"

  Ces phrases illustrent la polysémie du verbe "désigner". Il est donc nécessaire, le plus souvent, de recourir à des indices complémentaires pour lever l'ambiguïté des indicateurs verbaux, ce qui exige d'élaborer de nombreuses règles de reconnaissance pour un seul marqueur.

  Dans notre étude, nous supposons que le contexte pragmatique suffit à lever la polysémie d'un type pédagogique. En effet, nous nous limitons dans notre étude à des structures linguistiques souvent énoncés dans les documents pédagogiques.Si le contexte ne suffit pas, bien souvent, à lever la polysémie

des indicateurs d'un objet pédagogique de type bien défini, bien des énoncés peuvent être captés lorsque le travail de l'activité pédagogique laisse des traces (structures linguistiques discursives) d'une autre nature. Parmi ces marques, nous trouvons des indices nominaux, ainsi que des indices liés à la condition :

- Les indices nominaux

  Ces indices servent bien souvent à lever l'indétermination sur la valeur à donner à un indicateur verbal. On peut distinguer d'abord des nominaux qui expriment une signification associée à un point de vue de fouille à savoir : définition ; sens ; signification ; connotation. Par exemple : nous trouvons les synonymes ou termes proches de "mot" : terme ; expression ; vocable ; lexème ; dénomination ; appellation. Dans l'exemple précédant, la phrase "le symbole X désigne un des constituants de la phrase" reflète le point de vue "Définition" grâce au verbe "désigne" qui représente un indicateur principal.

  Dans les faits, lorsque nous construisons des règles d'exploration contextuelle, nous n'élaborons pas ces listes en fonction d'une proximité sémantique serrée, mais davantage sur la position de l'indice vis-à-vis d'un indicateur donné : on piochera indifféremment dans ces différentes listes pour construire des listes d'indices que l'on trouve à gauche de tel verbe, telle autre liste à droite de tel marqueur.

  Par exemple, dans la phrase "le symbole X désigne un des constituants de la phrase", en plus de l'indicateur "désigne", nous trouvons le mot "symbole" qui représente un indice à gauche levant la polysémie du verbe "désigner". Des synonymes du mot "symbole" peuvent aussi lever la polysémie du verbe "désigner".

- Les indices verbaux

  Ces indices verbaux organisent les verbes qui nous permettront de repérer des segments textuels exprimant des objets pédagogiques. Les verbes sont organisés dans des listes et sont de différentes natures :

  Les verbes qui traduisent les différentes formulations du type "Définition" sont : Définir, signifier, veut dire, etc. Pour le type "Caractéristique", les verbes sont : se caractériser, se distinguer, posséder, etc. les verbes qui traduisent le type "Exercice" sont : répondre, formuler, résoudre, etc. Il ya aussi les verbes auxiliaires comme "être", "avoir", etc. et les verbes modaux "pouvoir", "devoir",etc. Parmi ces verbes, il y a les verbes qui représentent des indicateurs principaux (verbes de type 1) et d'autres verbes qui ont comme fonction des indices auxiliaires (les verbes de type 2 et3).

– Les indices liés à la condition

Certains verbes susceptibles de porter un type d'objet pédagogique bien défini ont parfois besoin d'être repérés en présence d'indices qui marquent une condition, tels que "si", "si et seulement si", "quand", "lorsque". Par exemple : "Un quadrilatère est un parallélogramme si, et seulement si, ses diagonales ont le même milieu". L'expression de la condition peut en effet être une composante importante de la définition : nous retrouvons là l'idée qu'une définition repose sur l'établissement d'une condition nécessaire et suffisante.

- **Modélisation de la carte sémantique des types d'objets pédagogiques**

Plusieurs typages de l'objet pédagogique sont possibles selon le rôle ou l'usage potentiel de l'objet. Dans notre système, le critère est le rôle pédagogique, qui a pour but de susciter l'action pédagogique chez l'apprenant. Ainsi les objets pédagogiques sont caractérisés selon qu'ils sont : Définition, Exemple, Exercice, Plan, etc. Nous avons assigné ces différents rôles à des types d'objets pédagogiques.
Les différents types d'objets pédagogiques, que nous proposons, sont formulés comme suit :

- Le type "Définition"

Ce type d'objet a été étudié par Charles Teissedre dans son travail de mastère (Teissedre, 2007) où il le considère comme un point de vue. Dans le cadre de son mastère, il aborde le point de vue "Définition" avec une vision large. En fait, il étudie l'énoncé définitoire dans toutes ses formes pouvant être déposées dans n'importe quel type de document (articles de presse, articles philosophiques, livres, etc.). Notre travail est plus simple que le sien du fait que nous ne considérons que les définitions pouvant exister dans les documents pédagogiques (principalement les supports de cours).
D'après le dictionnaire Larousse, une définition est le "Fait de déterminer les caractéristiques d'un concept, d'un mot, d'un objet, etc., ensemble des propriétés essentielles de quelque chose.
Dans la "Logique", Larousse définit le mot "Définition" comme "Énoncé ou déclaration aux termes desquels un symbole ou une combinaison de symboles nouvellement introduits (definiendum) signifie ou dénote la même chose qu'un symbole dont le sens est déjà connu (definiens)."
Nous présentons ci-dessous quelques exemples de passages textuels exprimant le type "Définition" et que nous avons identifiés dans notre corpus :

4.1 Définition du génie logiciel (Software Engineering) :

- *Génie logiciel : élaboration et l'utilisation des principes de génie permettant de produire économiquement des logiciels fiables et qui fonctionnent de façon efficace sur des machines réelles.*

- *Génie logiciel : application pratique de la connaissance scientifique dans la conception et l'élaboration de programmes informatiques et de la documentation associée nécessaire pour les développer, les mettre en oeuvre et les maintenir (Bohem 1976).*

- *Génie logiciel : ensemble des activités de conception et de mise en oeuvre des produits et des procédures tendant à rationnaliser la production du logiciel et son suivi.*

Nous avons choisi de catégoriser le type "Définition" en 3 sous-types qui sont "Explication", "Signification" et "Formulation de condition". Nous présentons ci-dessous des exemples d'extraits textuels représentant chacun de ces sous-types :

- Le sous type "Explication" :

***Taux de fréquence TF** : C'est le nombre des cas de lésions par million d'heures de travail effectuées par toutes les personnes exposées au risque :*

$$TF = \frac{nombre\ total\ d'accident\ avec\ arrêt}{nombre\ d'heures\ travaillées} \times 10^6$$

- Le sous type "Signification" :

Un système d'information (SI) peut être considéré comme un ensemble de flux d'informations, d'opérations qu'ils subissent et de moyens mis en oeuvre pour ce faire quelque soit la nature de ces moyens.

- Le sous type "Formulation de condition"

Définition d'un système d'information automatisé : C'est une partie du système informatique regroupant uniquement les applications.

- Le type "Exercice"

  D'après le dictionnaire Larousse, l'Exercice est défini comme étant "une activité spécialement structurée, adaptée, qui permet de développer les capacités de quelqu'un dans un domaine : *Des exercices respiratoires.*". C'est l'action

de mettre de mettre en pratique une faculté, de faire valoir un droit : *L'exercice du pouvoir.* Dans un cadre pédagogique, c'est "un Problème, devoir, ensemble de questions dans lesquels on a à appliquer ce qui a été appris précédemment dans un cours". L'exercice représente l'action d'exercer ou de s'exercer (Wikipédia). Notre vision du terme "Exercice" est la suivante "C'est une procédure qui sert à renforcer des mécanismes des définitions, des exemples, des manipulations". L'exemple d'exercice suivant est identifié dans notre corpus :

Questions

1. *Faites un rapprochement entre Merise et UML ;*
2. *Faites un rapprochement entre Merise et le Génie Logiciel.*

Nous avons décomposé le type "Exercice" en trois sous types à savoir : "Application", "Questions à choix multiples" et "Etude de cas". Des exemples de segments textuels relatifs à ces sous-types sont donnés ci-dessous :

– Le sous type "Application"

Exercice 1

Considérez le réseau Ethernet suivant, composé d'un répéteur et de trois stations appelées ici DTE (Digital Terminal Equipment).
On souhaite déterminer, à l'aide des tables 1 et 2 fournies ci-dessous, si ce réseau qui est constitué d'un seul domaine de collision est viable ou non (s'il respecte ou non les règles de configuration pour les systèmes Ethernet à 10 mégabits, également connues sous le nom Transmission System Model).

Avant de se lancer dans les calculs nécessaires pour déterminer si ce réseau est conforme ou non à la norme, on prendra soin :

1. *de définir ce qu'est le délai d'aller-retour correspondant à un chemin entre deux DTE dans un réseau Ethernet (domaine de collision) ;*
2. *de rappeler ce qu'est une collision tardive en Ethernet et quand elle survient ;*

– Le sous type "Etude de cas"

EXERCICE 4 : Gestion des Travaux d'un Groupe de Recherche Objectif : *appliquer une démarche par étape pour l'élaboration du modèle. Nous voulons modéliser le système d'information relatif à la gestion des travaux d'un groupe de recherche. Ce groupe est constitué de chercheurs dont on connaît pour chacun le numéro, le nom, le prénom, le diplôme, l'activité de recherche, le responsable de recherche (lui-même un chercheur), l'adresse et le téléphone. Les chercheurs rédigent des articles dont chacun est caractérisé par un titre, le code et le titre du domaine de recherche, une date de rédaction et un certain nombre de mots clés qui ont pour rôle de faciliter la recherche documentaire. Un article peut être rédigé par plusieurs chercheurs. Nous supposons que le titre de l'article permet de l'identifier.*
Le groupe de recherche anime également des séminaires. Chacun est identifié par titre, le lieu et la date, et on détient aussi le responsable et les conférenciers qui font partie du groupe de recherche. Différents participants assistent aux séminaires. . . .
. . . . . .
Elaborez le modèle entité association associé à ce système d'information.

– Le sous type "Questions à choix multiple"

La question comporte un texte suivi d'une question sur ce texte :
Dans "Les misérables" Victor Hugo parle ainsi du policier Javert :
"Cet homme était composé de deux sentiments très simples et relativement très bons, mais qu'il faisait presque mauvais à force de les exagérer : le respect de l'autorité, la haine de la rébellion. Et à ses yeux, le vol, le meurtre, tous les crimes, n'étaient que des formes de rébellion. Il enveloppait dans une sorte de foi aveugle et profonde, tout ce qui a une fonction dans l'état, depuis le premier ministre jusqu'au garde champêtre".
Indiquez les affirmations correctes :

- *Javert ne respecte pas l'autorité*
- *Javert considère que les crimes sont des formes de rébellion*
- *Javert hait les représentants de l'état*
- *Javert admire tous les représentants de l'état, quel que soit leur grade*

- Le type "Exemple"

  D'après le dictionnaire Larousse, un exemple est défini comme étant "Ce qui peut servir de modèle, ce qui peut être imité".
  Un autre sens du mot "Exemple", celui de notre contexte est "Texte, fait illustrant des propos". Son synonyme est "Illustration". Ça peut être défini aussi comme un mot ou phrase qui illustre une définition, une règle. En pédagogie,

c'est l'étude d'une situation sur laquelle s'applique des résultats plus généraux énoncés avant. Nous présentons ci-dessous des exemples de segments textuels exprimant le type pédagogique "Exemple" :

Quelques exemples de modèles
***Modèle météorologique** : à partir de donn*
de prévoir les conditions climatiques pour le
***Modèle économique** : peut par exemple pe*
en fonction d'hypothèses macro-éc
de croissance...).
***Modèle démographique** : définit la compos*
comportement, dans le but de fiabiliser des é
de démarches commerciales, etc...

Dans notre corpus, nous avons identifié l'exemple suivant :

Le type "Exemple" est divisé en quatre sous-types : "Illustration", "Explication", "Contre-exemple". Nous présentons dans ce qui suit des exemples d'extraits textuels représentant chacun de ces sous-types :

- Le sous type "Illustration"
  Lorsqu'un exemple suit une idée dans un texte argumentatif, il l'éclaire, la précise. C'est un exemple illustratif.

EXEMPLES DE TAUX DE FRÉQUENCE

par comités techniques nationaux

- *Bâtiment et travaux public*	: 57,6
- *Bois, ameublement, papier carton*	: 35,1
- *Services, commerces et industrie de l'alimentation*	: 33,3
- *Métallurgie*	: 27,9
- *Transport, eau, gaz, électricité*	: 25,2
- *Chimie, caoutchouc, plasturgie*	: 23,2
- *Commerce non alimentaire*	: 14,5
- *Activités de service et travail temporaire*	: 27,6

- Le sous type "Argumentation" Si l'exemple précède l'idée, il présente un cas concret et permet de tirer un enseignement général, un argument ou une conclusion. On l'appelle exemple argumentatif.

Exemples :

- *Un jeune ouvrier met en route une perceuse, mais en oubliant de retirer la clé du mandrin porte-foret. La clé est projetée au loin, blesse l'ouvrier ou un de ses compagnons. L'accident est bien survenu à l'occasion de l'utilisation d'une machine, mais est dû en réalité à la négligence ou à l'ignorance.*
- *Un ouvrier, pour voir si une ligne 110 V est sous tension, la touche du doigt. Par malheur, son autre main est en contact avec tuyauterie métallique constituant un excellent conducteur de retour à la terre. L'ouvrier est électrocuté. L'accident est bien dû à l'électricité, mais il a été provoqué par une imprudence.*
- *Un mécanicien, pour réparer un tour parallèle, enlève le protecteur des engrenages de la tête de cheval. La réparation terminée, il omet de remettre le protecteur en place. Pendant qu'il essaie la machine, un de ses compagnons passe à proximité du tour et se fait happer le doigt. Là aussi, l'accident est le résultat d'une négligence grave.*

- Le sous type "Contre-exemple"
  Lorsqu'un exemple contredit une idée générale (c'est-à-dire soutient une thèse adverse), on l'appelle un **contre-exemple**.
  Des exemples qui vérifient l'énoncé ne suffisent pas à prouver que l'énoncé est vrai. Par contre un seul exemple qui ne vérifie pas l'énoncé suffit à prouver que l'énoncé est faux. Cet exemple est appelé un contre-exemple.

  Exemple : 18 est un multiple de 2 et ne se termine pas par 2 donc l'énoncé "tous les multiples de 2 se terminent par 2" est faux.

- Le type "Plan"

  Nous désignons par "Plan" les grandes lignes d'un document. Il peut être sous forme de table de matières d'un rapport, ou encore sous forme de sommaire d'un support pédagogique. Dans notre corpus, nous avons identifié des exemples de plans de sous types "Sommaire", "Table des matières", "Plan de cours". Un exemple du sous type "Plan de cours" est présenté dans l'annexe A (Document pédagogique n°2).

- Le type "Caractéristique"

  Par ce type, nous désignons ce qui caractérise, ou qui est un des traits dominants de quelque chose. C'est ce qui constitue la particularité, le caractère distinctif de quelqu'un ou de quelque chose. Dans notre contexte, nous considérons ce type comme les avantages, les inconvénients et les traits

distinctifs d'une notion bien déterminée. Nous illustrons, dans ce qui suit, des exemples de segments textuels représentant le type "Caractéristique" :
Un exemple ci-dessous d'objet pédagogique de type "Caractéristique" :

Un logiciel est caractérisé par :

- *Une diversité des applications : un logiciel est un produit atypique comparativement aux produits industriels : on peut parler d'avion type, de voiture type mais il est quasiment impossible de parler de logiciel type.*
- *Une taille et une complexité*
- *Abstraction et invisibilité*

Le type "Caractéristique" est divisé en trois sous-types à savoir "Points forts", "Points faibles", "Signes distinctifs".

- Le sous type "Points forts" :

Avantages du modèle en spirale

· *Le caractère itératif du modèle,*

· *La considération des risques pouvant être appréhendés dès le début du projet permettant ainsi de réviser les choix effectués au cours des différents cycles en fonction de l'avancement du projet,*

· *La rapidité de production d'un logiciel opérationnel même s'il est minimale,*

· *Il permet de se concentrer sur les aspects les plus incertains du développement,*

· *Il tolère la remise en cause de la part du client à chaque nouvelle évaluation*

- Le sous type "Points faibles"

Inconvénients du modèle en spirale

· *Risque de remise en cause des spécifications des versions déjà réalisées lors de l'analyse de nouvelles versions,*

· *Difficultés de mise en œuvre au niveau procédural et de contrôle du processus,*

· *Organisation opérationnelle du développement souvent modifiée pour le client,*

· *Difficultés pour mener à bien les premiers cycles de la spirale.*

- Le sous type "Signes distinctifs"

***D'après la cour de cassation** : L'accident de travail est légalement caractérisé par **l'occasion violante** et **soudaine** d'une cause extérieure provocant, au cours du travail, **une lésion corporelle**.*

- Le type "Cours"

  D'après (Michel et *al.*, 2002), le cours est l'ensemble des ressources choisies pour présenter une matière ou un savoir. Il est défini soit par des objectifs d'enseignement (ou d'apprentissage) ayant une finalité précise, soit par un ensemble de connaissances que l'étudiant doit acquérir. Il est décrit par :

  – le titre du cours, sa description, une finalité,
  – un ensemble d'objectifs dans lesquels l'enseignant exprime son intention pédagogique c'est-à-dire l'ensemble des changements durables qu'il souhaite voir se produire chez l'apprenant,
  – un ensemble de thèmes (un thème correspond à un élément du contenu de la matière dont la maîtrise passe par la réalisation d'un ou de plusieurs objectifs pédagogiques).

  Nous considérons le type "Cours" comme un type représentant tout support de cours utilisé pour l'enseignement ou l'apprentissage. L'objet pédagogique de type "Cours" est généralement composé de plusieurs autres objets de différents types, comme "la définition", "l'exercice", "l'exemple", "la synthèse", etc. Il peut réunir au moins deux objets de types présentés ci-dessus.
  Dans notre corpus, nous avons identifié plusieurs objets de type "Cours". Un exemple de ce type d'objet est présenté dans l'annexe A(Document pédagogique n°2).

- Le type "Méthode"

  D'après le dictionnaire Larousse, une méthode est définie comme :

  - Marche rationnelle de l'esprit pour arriver à la connaissance ou à la démonstration d'une vérité : *La méthode se différencie de la théorie.*
  - Ensemble ordonné de manière logique de principes, de règles, d'étapes, qui constitue un moyen pour parvenir à un résultat : *Méthode scientifique.*
  - Manière de mener, selon une démarche raisonnée, une action, un travail, une activité ; technique *Une méthode de travail. Les méthodes de vente. Il n'a suivi aucune méthode précise dans son enquête.*
  - Ensemble des règles qui permettent l'apprentissage d'une technique, d'une science ; ouvrage qui les contient, les applique : *Méthode de lecture.*

  Nous présentons ci-dessous un exemple d'objet pédagogique de type méthode que nous avons identifiée dans notre corpus.

L'essai de traction (défini dans la norme NF EN 10002) consiste à soumettre une éprouvette à un effort de traction, et cela généralement jusqu'à rupture en vue de déterminer une ou plusieurs caractéristiques mécaniques.

- Relations liant les différents types d'objets

  Après avoir défini les différents types d'objets, nous exposons les relations éventuelles qui peuvent exister entre ces différents types d'objets en dedans d'un document pédagogique. Ces relations seront présentées ci-dessous :

  - La relation "Exemple-peut-être-Définition"
    Nous présentons ci-dessous un exemple d'un objet pédagogique de type "Exemple" et qui s'agit en même temps d'une "Définition".

Voici un exemple de définition du terme "Génie Logiciel" : Le GL est l'élaboration et l'utilisation des principes de génie permettant de produire économiquement des logiciels fiables et qui fonctionnent de façon efficace sur des machines réelles.

  - La relation "Exemple-peut-être-Exercice" :

Exemple *: Une entreprise occupe 500 personnes, qui travaillent chacune 50 semaines par an et 48 heures par semaine. Le nombre des cas de lésion professionnelle, au cours d'une année, a été de 60. Le nombre de journée chaumé en conséquence a été de 528. Pour cause de maladies ou d'accidents et pour d'autres raisons, les travailleurs ont été absents pendant 5 pour cent du nombre total possible d'heures de travail. Déterminer :*

· *Le taux de fréquence*

· *L'indice de fréquence*

· *et le taux de gravité dans cette entreprise durant l'année considérée.*

  - La relation "Exemple-peut-être-Caractéristique" :
    Nous présentons ci-dessous un exemple d'un objet pédagogique de type "Exemple" et où il s'agit en même temps d'une "Caractéristique".

Voici un exemple d'une caractéristique d'un logiciel : Sa taille et sa complexité.

A partir de cette modélisation et des différentes considérations établies précédemment, il est possible de regrouper et classer, dans une carte sémantique, les énoncés relatifs à chaque type d'objet pédagogique selon différents critères et en différentes sous catégories. Chaque nœud de la carte sémantique que nous présentons ici est un concept qui reçoit une étiquette sémantique avec laquelle nous pouvons annoter les segments textuels relatifs à chacun des types d'objets pédagogiques. Chaque concept est inséré dans une structure (celle de la carte).

Cette carte est le résultat d'un certain nombre de choix. En effet, par exemple, "La signification est-elle une sous classe de la définition ?". Ce choix de présentation est contestable, mais s'explique par le fait que, dans l'implémentation informatique des ressources linguistiques, le type d'objet général à rechercher est celui de l'Exemple, Exercice, Définition, etc. : toute annotation doit donc appartenir à l'un de ces types.

La carte sémantique des types d'objets pédagogiques, présentée sous forme d'un graphe(cf. Fig.3.3), a pour objectif de préparer l'implémentation des ressources linguistiques que nous avons rassemblées : elle est destinée à orienter cette implémentation, ainsi qu'à permettre un regroupement des marques linguistiques (selon chaque type). La constitution de règles de reconnaissance et d'extraction des différents types d'objets est donc censée découler de cette carte : dans les faits toutefois, la carte a été établie au fur et à mesure, par un long travail qui a accompagné plutôt que précédé le travail d'implémentation. La clarté de la présentation exigeait ce découpage entre l'analyse linguistique et l'automatisation, mais l'un et l'autre ont été menées de front par une série de va et vient en spirale qui s'enrichie, et dont les résultats sont en permanence confrontés (Smine et *al.*, 2010 (a), (b)).

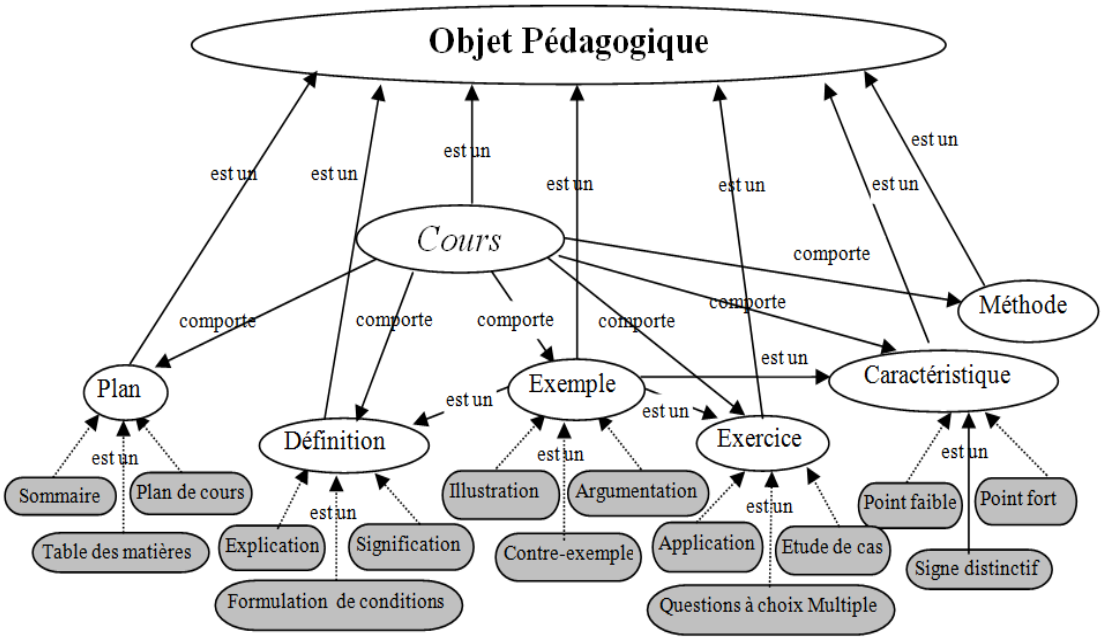


FIGURE 3.3: Carte sémantique des différents types d'objets pédagogiques

- **Annotation selon les différents types d'objets dans les textes**

Ce module est détaillé en plusieurs sous étapes et fait l'objet de plusieurs publications scientifiques comme (Smine et *al.*, 2011(a),(b)). Les sous étapes sont présentées ci-dessous :

- Repérage des indicateurs

  Un repérage des indicateurs est effectué pour chacune des règles relatives à chaque type d'objet pédagogique. En fait, dans les différents segments constituant chaque document pédagogique, une recherche des indicateurs potentiels est lancée.

  Si le système identifie au moins un indicateur dans l'espace de recherche (le segment dans notre cas). Par exemple, s'il a identifié l'indicateur "un exemple sur", il passe à l'étape suivante du processus d'annotation (Sélection des règles candidates). En effet, généralement les indicateurs n'expriment pas le type d'objet recherché. La présence d'indices linguistiques est primordiale pour le fonctionnement de la règle d'Exploration Contextuelle (Desclés et *al.*, 2007). Dans le cas où aucun indicateur de la présente règle n'a été identifié, le système prend la règle suivante relative au même point de vue pour l'appliquer sur les documents.

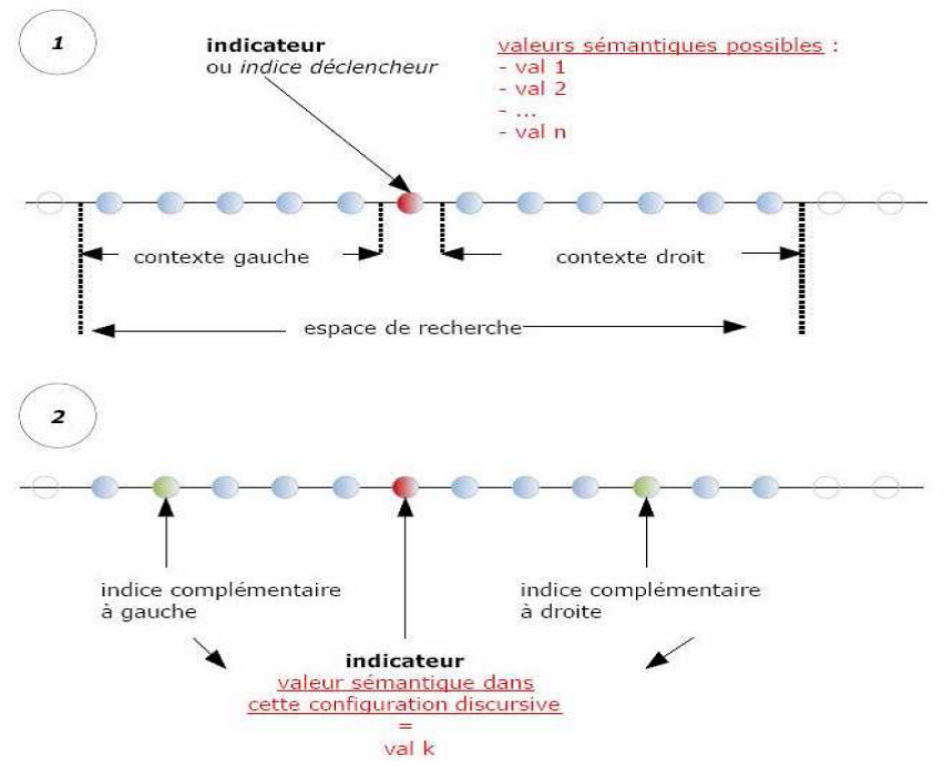


FIGURE 3.4: Schéma de fonctionnement d'une règle d'exploration contextuelle (1) (Desclés et *al.*, 2007)

- Sélection des règles candidates

  L'identification des indicateurs permet de retenir que les règles qui doivent être déclenchées. Ces règles sont nommées règles candidates. Le système passe à la vérification de la présence ou l'absence des indices relatifs à chaque indicateur. Une décision est par la suite prise quant à l'annotation du segment : Au cas où les indices sont vérifiés, la règle d'Exploration Contextuelle en question devient "applicable" pour annoter le segment en cours. Sinon, la procédure d'annotation par la règle d'Exploration Contextuelle en cours est annulée. L'espace de recherche de l'un de ces indices est à "Gauche" de l'indicateur. Si le système arrive à identifier, par exemple, l'indice "Nous prenons", alors une annotation est attribuée au segment en question portant comme valeur (étiquette) le point de vue "Exemple".

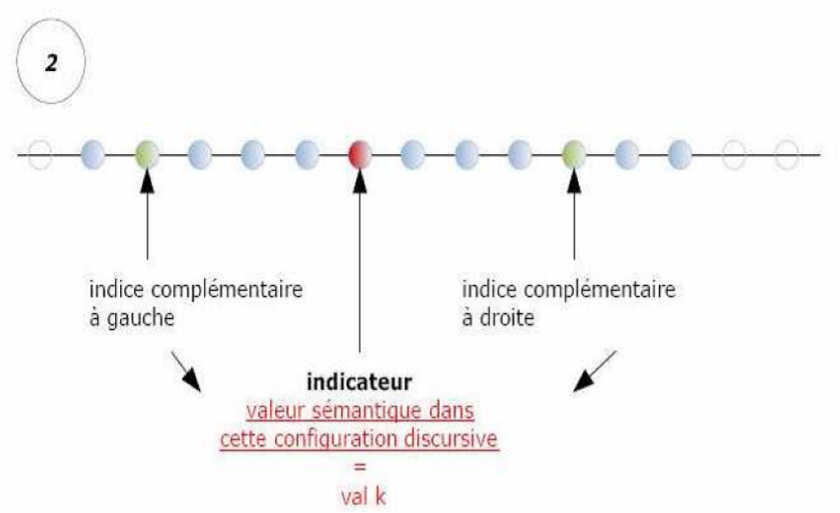


FIGURE 3.5: Schéma de fonctionnement d'une règle d'EC (2) (Desclés et *al.*, 2007)

Au cas où aucun indice n'a été repéré, l'application de cette règle est annulée.

Pour chaque annotation attribuée, le segment annoté est stocké dans une base de données Excel avec un ensemble d'informations qui lui sontrelatives : l'emplacement du document qui contient le segment en question, le type d'objet pédagogique annotant le segment (Définition, Exemple, Exercice, cours, ...) et le titre du paragraphe annoté. Cet ensemble d'informations servira à l'évaluation du module d'annotation.

Pour chaque catégorie de la carte sémantique (cf. Fig.3.3), nous avons

défini l'ensemble des règles couvrant toutes les formes linguistiques possibles relatives aux objets pédagogiques. Nous avons développé environ 100 règles. Nous commençons par un exemple textuel pour généraliser ensuite toutes les structures linguistiques. Cette méthode permet de définir d'une manière incrémentale une base de règles solide.

Ces règles, dont le fonctionnement est décrit plus haut, permettent de repérer la valeur sémantique recherchée et qui doit être relative à un type d'objet pédagogique.

Ci-dessous des exemples de règles écrits en langage formel :

La règle RD2

***Identifiant de la Règle** : RD2*
***Type pédagogique** : Définition*
***Liste des indicateurs I** : défini, définie, définies, définis*
***Liste des indices gauches CG1** : est, était, sont, étaient*
***Liste des indices droite CD1** : comme*
***Condition 1** : il existe une occurrence d'un indicateur appartenant à la liste des indicateurs I*
***Condition 2** : il existe une occurrence d'un indice à gauche de l'indicateur appartenant à la liste des indices CG1 et une occurrence d'un indice à droite de l'indicateur appartenant à la liste des indices CD1.*
***Action** : annoter le segment textuel (la phrase) par le type pédagogique Définition*

FIGURE 3.6: Exemple de règle RD2

Cette règle suit ces étapes pour annoter un segment textuel (une phrase) comme une Définition :

- Exprimer la sémantique de la catégorie "Définition"à travers un indicateur parmi les verbes suivants : défini, définie, définies, définis, etc.
- Pour confirmer la valeur sémantique de l'indicateur du type pédagogique Définition, le système doit identifier, en premier lieu, dans la phrase des termes de la liste CG1 (est ou était ou sont ou étaient) dans le contexte gauche de l'indicateur.
- L'indicateur a besoin d'une autre expression comme la préposition comme dans le contexte droit pour permettre l'annotation de la phrase comme une Définition.

Nous remarquons que l'indicateur peut prendre plusieurs valeurs, séparée chacune par une virgule. Elles correspondent ici à un certain nombre de flexions du verbe 'définir'. Si l'une des valeurs de l'indicateur est identifiée dans une phrase (l'espace de recherche), alors l'Exploration Contextuelle est

déclenchée. Un premier indice est d'abord recherché à gauche de l'indicateur (contexte gauche). Si cet indice est présent, un second indice est alors recherché à droite de l'indicateur (contexte droit) : ici, la préposition comme. S'il est repéré, la phrase est annotée par l'étiquette Définition.
L'exemple ci-dessous de phrase annotée par la règle RD2 répondant à la structure recherchée :

La maintenance est définie comme l'ensemble des activités destinés à maintenir ou à rétablir un bien dans un état de sûreté de fonctionnement.

L'indicateur repéré ici est "définie". Les indices qui ont corroboré la valeur définitoire de l'indicateur sont "est" et "comme" ; ils sont situés respectivement à gauche et à droite de l'indicateur.
L'annotation a pour valeur "Définition", soit le premier niveau de la carte sémantique.
Les règles d'Exploration Contextuelle nous font donc entrer dans le tissu discursif et sémantique de la phrase, puisque le repérage d'un certain type de structure, mis à jour par un travail linguistique, permet de signaler que l'énoncé relève, dans ce cas, d'une définition. Il s'agit de reproduire automatiquement la méthode de lecture et d'extraction d'une personne qui chercherait à repérer une information dans un texte : elle s'arrêterait à quelques marques saillantes, avant de voir si d'autres marques plus ténues viennent confirmer qu'elle est face à l'information recherchée.

La règle RCO13

***Identifiant de la Règle** : RCO13*
***Type pédagogique** : Cours*
***Liste des indicateurs I** : Cours|cours|Chapitre|chapitre|Support de cours| Supports de cours|support de cours|supports de cours|Document|document|Documents|documents*
***Condition 1** : il existe une occurrence d'un indicateur appartenant à la liste des indicateurs I au niveau du titre*
***Action** : annoter le segment textuel (le document) par le type pédagogique Cours*

FIGURE 3.7: Exemple de règle RCO13

Cette règle suit ces étapes pour annoter un segment textuel (un document) comme un Cours :

- Exprimer la sémantique du type pédagogique "Cours" à travers un indicateur parmi les noms suivants : *Cours, cours, Chapitre, chapitre, Support de cours, Supports de cours, support de cours, supports de cours, Document, document, Documents, documents — etc.*

***Identifiant de la Règle** : RC5*
***Type pédagogique** : Caractéristique*
***Liste des indicateurs I** : caractéristique|caractéristiques*
***Liste des indices gauches CG1** : La| les|une| des| Les|Des|Une*
Liste des indices droite CD1 : de|des|du ***Condition 1** : il existe une occurrence d'un indicateur appartenant à la liste des indicateurs I dans un segment textuel (une phrase)*
***Condition 2** : il existe une occurrence d'un indice à gauche de l'indicateur appartenant à la liste des indices CG1 et une occurrence d'un indice à droite de l'indicateur appartenant à la liste des indices CD1.*
***Action** : annoter le segment textuel par le type pédagogique Caractéristique*

FIGURE 3.8: Exemple de Règle RC5

- Pour confirmer la valeur sémantique de l'indicateur du type pédagogique Cours, le système doit identifier cet indicateur dans le titre

Nous remarquons que l'indicateur peut prendre plusieurs valeurs, séparée chacune par une virgule. Elles correspondent ici à un certain nombre de flexions du nom "Cours". Si l'une des valeurs de l'indicateur est identifiée dans le titre, alors le document st annoté par le type *Cours*.

La règle RC5

Cette règle RC5 suit ces étapes pour annoter un segment textuel (une phrase) comme une Caractéristique :

- Exprimer la sémantique du type "Caractéristique" à travers un indicateur parmi les noms "caractéristique" et "caractéristiques".
- Pour confirmer la valeur sémantique de l'indicateur du type pédagogique Caractéristique, le système doit identifier, en premier lieu, dans la phrase des termes de la liste CG1 (voir règle) dans le contexte gauche de l'indicateur.
- L'indicateur a besoin d'une autre expression (du, de, des) dans le contexte droit pour permettre l'annotation de la phrase comme une Caractéristique.
    - L'emplacement du terme de la requête

Nous avons ajouté un composant à chaque règle qui représente l'emplacement du terme de la requête à rechercher dans le cadre du segment textuel exprimant l'objet pédagogique (Smine et *al.*, 2011(c)). Le besoin d'ajouter ce composant est né de la variation de l'emplacement du terme à rechercher avec la variation des structures langagières exprimant les objets pédagogiques. Ceci permet d'identifier les segments textuels exprimant le type d'objet pédagogique ainsi que le concept demandé par l'utilisateur. Par exemple, pour le même type d'objet pédagogique "Définition" : le terme à

rechercher "*Maintenance*" peut se trouver au début du segment "*La maintenance est définie comme l'ensemble des activités destinés à maintenir ou à rétablir un bien dans un état de sûreté de fonctionnement*" ou au milieu du segment pour le cas "*L'AFNOR a défini la maintenance comme étant l'ensemble des activités de remise en état de fonctionnement d'un système*". Sans la considération de ce paramètre, le système peut ne pas extraire l'objet demandé par l'utilisateur comme, pour le type Cours, où la plupart des règles d'EC exigent un emplacement du terme de la requête au niveau du Titre du document. Au cas où le terme est recherché hors du titre, le résultat de la recherche sera erroné.

De ce fait, l'emplacement du terme est un paramètre qui diffère d'une règle à une autre, selon la structure langagière exprimée par cette dernière. Nous avons désigné cet emplacement par une étiquette, qui prendra une valeur parmi un ensemble fini de valeurs désignant l'emplacement du terme par rapport aux indicateurs et indices (cf. Tab.3.3). Par exemple, GIND indique le terme, et se place à gauche de l'indicateur, ou TITRE indique que l'emplacement du terme est au niveau du titre du document. En fait, dans plusieurs cas, le titre peut nous révéler des connaissances sur le contenu du document.

Désignation de l'Emplacement du terme	**Emplacement du terme**	**Exemple**
TITRE	Dans le titre du document	
Gindicateur	A gauche de l'indicateur	
Dindicateur	A droite de l'indicateur	
DCD1	A droite du premier indice à droite	
GCD1	A gauche du premier indice à droite	
GCD2	A gauche du deuxième indice à droite	
GCG1	A gauche du premier indice à gauche	
GCG2	A gauche du deuxième indice à gauche	
DCG2	A droite du deuxième indice à gauche	

Tableau 3.2: Différentes possibilités de l'emplacement du terme de la rquête

Comme nous l'avons déjà mentionné, l'ensemble des règles est stocké dans un fichier Excel. Dans le cadre de notre méthode, nous donnons à l'utilisateur la possibilité de gérer la base de ses règles en ajoutant, supprimant ou modifiant des règles. Les informations associées à chaque règle son représentées dans le

tableau suivant (Nous prenons l'exemple de la règle RD2 (cf. Fig.3.6) pour illustrer une application de ces informations :

Nom du champ	Désignation	Exemple
IDR	Identifiant de la règle	RD2
Type	Le type pédagogique de la règle	Définition
Sous-type	Le sous-type pédagogique de la règle	Explication
Indicateur	La liste des indicateurs qui déclenchent cette règle	défini|définies|définie| définis|considéré|considérée| considérée|considérées
CG1	La première liste des indices à gauche de l'indicateur	est|était|a été|sont|ont été
CG2	La deuxième liste des indices à gauche de l'indicateur	
CGN	La liste des indices négatifs à gauche de l'indicateur	
CD1	La première liste des indices à droite de l'indicateur	par|comme
CD2	La deuxième liste des indices à droite de l'indicateur	
CDN	La liste des indices négatifs à droite de l'indicateur	
EmpTerme	L'emplacement du terme de la requête par rapport à l'indicateur et aux indices	GCG1 (A gauche de l'indice gauche)

Tableau 3.3: La liste des informations associées à chaque règle

Nous prenons un extrait de texte à partir d'un document pédagogique

Cours sur l'informatisation des systèmes d'information
2.Informatisation
L'informatisation est généralement définie comme la mise en place d'un système de traitement automatique de l'information dans un service n'utilisant pas l'informatique au préalable.
Le processus d'informatisation des systèmes informatiques comprend deux activités principales : activité de développement et activité de maintenance.
Avantages :

- *Meilleure productivité,*
- *Mondialisation,*
- *Rapidité de traitement,*
- *Coût de production faible.*

Inconvénients :

- Coût de développement t de maintenance élevé.

La règle RD2, appliquée à l'exemple ci-dessus, permet d'annoter la phrase *"L'informatisation est généralement définie comme la mise en place d'un système de traitement automatique de l'information dans un service n'utilisant pas l'informatique au préalable"* comme une Définition. Ce type d'objet est détecté grâce à l'expression "définie" qui est une occurrence Ii de l'indicateur du type de la règle RD2. Ensuite, la présence de l'indice gauche CG1 "est" et l'indice droite CD1 "comme" permet d'annoter le segment comme une Définition.
Pour le type "Cours", le repérage de l'occurrence Ii au niveau du titre est suffisant pour annoter le document comme un cours. L'indicateur nominal de l'objet pédagogique est le mot "Cours", et d'autres noms comme "Chapitre", "Notes de cours". A part le titre, l'existence de l'indicateur "Cours" n'implique pas l'annotation du document comme un cours.
Afin d'annoter le segment suivant comme une "Caractéristique" et plus précisément "points forts", nous devons détecter le mot "Avantages".

Avantages :

- *Meilleure productivité,*
- *Mondialisation,*
- *Rapidité de traitement,*
- *Coût de production faible.*

De même pour le segment suivant, il suffit de détecter le mot "*Inconvénients*" pour annoter le segment comme une "*Caractéristique*" et plus précisément "*Points faibles*".

"Inconvénients :- Coût de développement et de maintenance élevé."

Les informations introduites (annotations) à ces segments annotés sont les suivantes :

- Identifiant de la règle appliquée (IDR)
- Type de l'objet annoté (Type)
- Sous-type de l'objet annoté (SousType)
- Emplacement du terme de la requête (EmpTerme)

Ces informations sont introduites sous la forme de balises XML en-dessous du segment annoté.

```
<article>
 <section>
  <paragraphe>
   <phrase>
     <texte> texte 1 </texte>
     <annotation>
        <title> title 1 </title>
        <idr> id regle <idr>
        <empterme> emplacement terme </empterme>
        <type> le type </type>
        <soustype>le sous type</soustype>
        <texte>texte de l'annotation</texte>
     </annotation>
   </phrase>
  </paragraphe>
 </section>
</article>
```

3.5.2 Représentation vectorielle des objets pédagogiques

Suite à l'annotation, nous avons développé une direction de travail relativement indépendante du traitement sémantique (Exploration Contextuelle) de la langue naturelle, mais davantage liée aux statistiques et à la recherche documentaire. Elle partait plutôt des nécessités de la classification et recherche de documents (Salton et *al.*, 1983), (Salton et al., 1994). Cette direction numérique est plus proche des mathématiques, et en particulier des probabilités. Plutôt que de construire des structures langagières, nous utilisons le "modèle vectoriel" pour représenter les objets pédagogiques par des vecteurs. Pour donner un exemple, une application typique consiste à représenter des documents

par des vecteurs calculés à partir de mots les plus significatifs présents dans chaque document. Ces vecteurs sont ensuite appariés par rapport au vecteur de la requête utilisateur. Cette classification peut alors servir à l'indexation et à la recherche des documents, mais aussi à l'extraction d'objets pédagogiques.
Le modèle vectoriel, que nous avons déjà présenté dans le deuxième chapitre, est donc fondamental pour la représentation vectorielle des documents.
Nous allons dans ce qui suit présenter en détail les étapes de notre modèle de représentation vectorielle des objets pédagogiques.
Suite à l'annotation des objets pédagogiques selon leurs types, nous appliquons la méthode vectorielle de Salton (Salton et *al.*, 1975) pour créer des vecteurs représentatifs de ces objets annotés.
Notre choix est porté sur cette méthode pour les raisons suivantes : (1) l'algorithme de représentation vectorielle prend en compte le poids d'un terme basé sur l'occurrence de ce terme dans l'objet pédagogique, (2) Le processus habituel de recherche documentaire dans le modèle vectoriel représente la requête par un vecteur dans le même espace que les documents et compare ce vecteur à tous ceux de la matrice. Cette comparaison équivaut au calcul d'une fonction de similarité (ou de distance) entre les vecteurs représentant les documents et le vecteur correspondant à la requête. Elle permet d'ordonner les documents en fonction de leur ressemblance avec la requête, (3) l'algorithme de représentation vectorielle n'assigne pas une seule classe au document mais calcul une mesure de similarité entre la requête utilisateur et les différents objets pédagogiques, ce qui correspond à notre but dans cette étape.
Ce modèle a notamment été critiqué à cause de l'hypothèse d'indépendance des mots-clé (la dimension de l'espace correspond au nombre de mots-clés). Cependant, malgré sa simplicité apparente, le modèle vectoriel s'est au moins montré aussi bon (autant pour la qualité des résultats que pour la rapidité avec laquelle ils sont obtenus) que les autres modèles.

Les différentes étapes de la représentation vectorielle sont détaillées dans ce qui suit :

- **Prétraitement des objets pédagogiques** :

  Pour les objets pédagogiques, nous procédons à une extraction des termes en appliquant les outils de TALN. Contrairement à certains outils qui ignorent les mots qui appartiennent à une liste prédéfinie de mots vides et ne prennent en considération que les mots pleins, nous prenons en compte tous les termes composant un objet pédagogique. L'ensemble des termes représentent les termes index de chaque objet pédagogique. Signalons que nous employons souvent la notion de terme pour désigner un "mot" plus abstrait commun à toute une famille de mots.

	L'informatisation	est	l'installation	d'un	système	de	traitement	automatique	l'information	la	mise	en	place	dans	un	service	n'utilisant	pas	l'informatique	au	préalable
Obj1	1	1	0	1	1	2	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Obj2	1	1	1	1	1	2	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0

TABLEAU 3.4: Tableau Représentation vectorielle des deux objets

- **Transformation des objets pédagogiques en vecteurs de poids** :

  Dans cette étape, un objet pédagogique est représenté par un vecteur :
  $V^m = (p_1^m, p_2^m, ..., p_i^m)$ où p_i^m est le poids du terme index i dans l'objet m.
  l est le nombre total de termes de l'index.
  La pondération traduit la fréquence du terme index dans l'objet pédagogique. Nous avons choisi d'appliquer cette mesure de pondération car, dans notre cas, la taille d'un objet pédagogique n'est pas énorme (en moyenne 50 mots) ce qui rend inutile d'appliquer des mesures de pondération intelligentes. Nous présentons ci-dessous, en exemple, deux objets pédagogiques représentés par un vecteur de poids des termes index.

  Soit deux exemples d'objets pédagogiques, l'un extrait de l'exemple présenté précédemment ("*L'informatisation est la mise en place d'un système de traitement automatique de l'information dans un service n'utilisant pas l'informatique au préalable*") et l'autre que nous avons-nous même créé, à savoir : "L'informatisation est l'installation d'un système de traitement automatique de l'information".

Les vecteurs représentant les deux objets sont donnés ci-dessous :

3.5.3 Indexation sémantique des objets pédagogiques

L'indexation se situe dans un contexte plus global, qui est celui de l'analyse du contenu. Cette étape est un préalable indispensable à toute recherche d'information sur le contenu et à d'autres types de traitement des informations. Dans cette étape, différentes techniques sont proposées pour identifier et pondérer les termes les plus aptes à décrire le contenu des documents. La finalité de l'indexation est de permettre une recherche

efficace des informations contenues dans une collection de documents sans avoir à analyser chaque texte de document à chaque interrogation ou recherche (Goker et *al.*, 2009). Pour notre part, l'étape d'indexation réunit les résultats des deux précédentes étapes à savoir : l'annotation sémantique et la représentation vectorielle des objets pédagogiques.

3.5.3.1 Indexation des annotations des objets

L'équipe LaLIC (Djioua et al., 2006), (Djioua et al., 2007) développe une approche différente, qui se traduit dans l'élaboration de l'outil d'indexation Mocxe. La différence par rapport aux autres méthodes réside dans le fait qu'il ne s'agit pas d'utiliser des mots-clés pour refléter le contenu d'un texte et pour le retrouver : Mocxe s'appuie entièrement sur les résultats de l'annotation. Il propose ainsi un nouveau paradigme d'indexation, dont l'idée phare consiste à recourir à des "segments-clés" plutôt qu'à des "mots-clés" pour indexer les documents.
La stratégie retenue se déploie en deux étapes : la première consiste à annoter les documents textuels en fonction de différents points de vue sémantiques et discursifs et à les stocker ; la seconde étape consiste à indexer les segments annotés (les phrases et paragraphes), afin de produire des réponses qui ne se présentent pas uniquement sous la forme d'une liste de documents, mais également sous la forme de portions de textes pertinentes compte-tenu de la requête. Cette requête porte sur un point de vue sémantique donné par l'utilisateur (définition, causalité, connexion, citation, ...).
Nous nous sommes inspirés de ce travail pour développer notre méthode d'indexation des objets pédagogiques : Suite à l'annotation des objets pédagogiques selon leurs types, nous procédons à une indexation de ces objets afin de fournir à l'utilisateur les objets pertinents répondant à sa requête, non seulement sous forme de documents, mais aussi sous forme d'objets pédagogiques textuels.
L'étape d'indexation prend en entrée les objets pédagogiques annotés et les informations relatives à cette annotation à savoir :

- Le type de l'objet pédagogique annoté (Définition, Exemple, Exercice, etc.).
- Le sous-type de l'objet annoté
- Le chemin du document pour l'identifier
- L'identifiant de la règle appliquée pour l'annotation de l'objet
- L'emplacement du terme de la requête
- Le contenu textuel de l'objet pédagogique

Le résultat de l'indexation sera sauvegardé dans un fichier index sous la forme d'un ensemble d'informations sur chaque objet indexé. Nous donnons ci-dessous un exemple d'un objet indexé :
Soit l'exemple de l'objet pédagogique présenté précédemment à indexer :
"L'informatisation est la mise en place d'un système de traitement automatique de l'information dans un service n'utilisant pas l'informatique au préalable". Suite à l'annotation de cet objet, les informations suivantes sont introduites :

- Type : Définition
- Sous-type : Explication
- Chemin du document
- ID règle : RD1
- Emplacement terme : GIND
- Contenu de l'objet : "L'informatisation est la mise en place d'un système de traitement automatique de l'information dans un service n'utilisant pas l'informatique au préalable"

Les informations relatives à cet objet annoté seront ajoutées dans le fichier index.

3.5.3.2 Indexation des vecteurs représentatifs des objets pédagogiques

La représentation des documents dans le modèle vectoriel conduit à autant de vecteurs que de documents. La taille de ces vecteurs est égale au nombre de termes (mots clé) différents contenus dans le corpus à indexer.
L'utilisation d'un fichier inversé permet de diminuer considérablement les besoins en espace mémoire et le temps requis pour une recherche.
Un fichier inversé contient un vecteur (dictionnaire dont les composantes désignent les termes de l'index) et autant de listes de documents que de termes dans l'index (cf. Fig.3.9). À chaque terme de l'index (chaque composante), est en effet associée la liste des documents où il apparaît (chaque référence à un document est accompagnée du poids du terme dans ce document). De cette manière, la taille de l'index diminue fortement (elle représente environ le tiers de la taille du corpus) et la recherche est très rapide. Nous présentons ci-dessous un exemple de documents indexés (cf. Fig.3.9) : le terme T2 se trouve dans les documents 1,7,9 et 10.

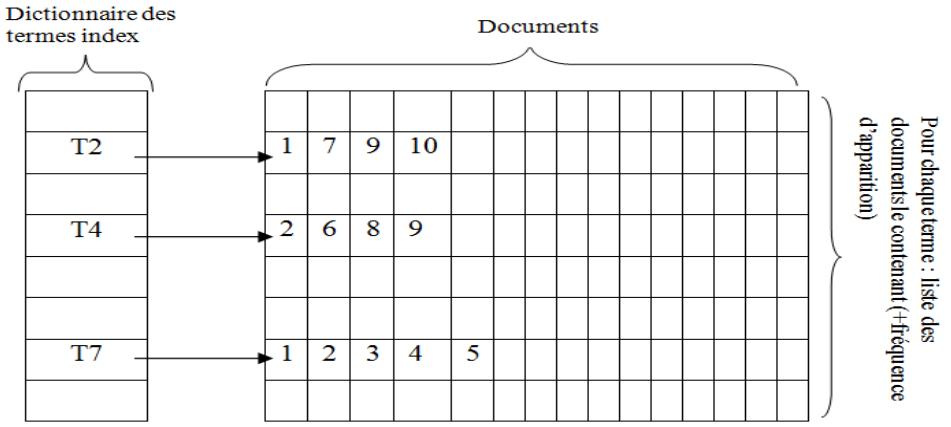


FIGURE 3.9: Exemple de fichier index

3.5.3.3 L'organisation du fichier index

Un fichier inversé est un index lexicographique, c'est-à-dire une table alphabétique de mots-clés accompagnés de références. Il permet à partir d'un mot-clé donné de trouver toutes ses occurrences au sein d'une collection de documents. Dans le cas général, il comporte, pour chaque terme d'indexation, une liste (appelée < posting list > ou parfois < posting >) contenant l'identifiant des documents dans lesquels il apparaît ainsi que sa fréquence d'apparition intra-document. Dans le cas où le fichier inversé mémorise en plus toutes les positions de chaque occurrence, le fichier inversé est dit < complet> (< full inverted file >) (Gillard, 2002).

L'avantage de cette structure est qu'elle permet de représenter, avec efficacité, l'ensemble de la collection des documents. Ainsi, en conservant une seule occurrence de chacun des termes d'indexation, elle diminue l'espace mémoire nécessaire. Enfin, elle accélère la recherche car elle supprime tout besoin d'accès aux documents d'origine : le fichier inversé contient toutes les informations utiles et la plupart des calculs numériques peuvent être effectués au moment de l'indexation.

Il existe plusieurs méthodes pour implémenter un fichier inversé. Elles peuvent être fondées sur l'utilisation d'un tableau (Harman et al., 1992) avec notamment un algorithme de construction à base de tableaux triés) ou d'une table de hashage. D'autres structures adaptées à la création d'index existent dans la littérature voir par exemple (Christos, 1992) pour les fichiers de signatures et (Gonnet et al., 1992) pour les tableaux et d'arbres PAT.

Dans le cadre de notre travail, nous avons eu recours à un fichier inversé à base de table de hashage. Le moteur d'indexation utilise une structure composée des objets pédagogiques d'un côté, leurs types pédagogiques de l'autre côté. Le moteur d'indexation n'utilise pas seulement les termes linguistiques à rechercher, mais explore aussi les annotations ajoutées aux différents objets d'un document. Nous rappelons que ces annotations se sont basées sur une étude des marqueurs linguistiques pour rechercher les structures linguistiques exprimant le type d'objet à annoter. L'organisation du fichier index de notre système présente la relation entre les documents textuels, les objets pédagogiques constituant le document et leurs types pédagogiques. Un document contient des objets pédagogiques identifiés par le système d'annotation d'une part et représenté par un vecteur de poids d'une autre part (Smine et *al.*, 2011(d), (e)), (Smine et *al.*, 2012).

Chaque objet pédagogique est alors associé avec plusieurs informations qui sont :

- Le type de l'objet pédagogique indexé (Définition, Exemple, Exercice, etc.).
- Le sous-type pédagogique de l'objet indexé
- Le chemin du document contenant l'objet
- L'identifiant de la règle appliquée pour l'annotation de l'objet
- L'emplacement du terme de la requête
- Le contenu textuel de l'objet pédagogique
- Le vecteur de poids représentant l'objet pédagogique

L'organisation de l'index de notre système est présentée dans la figure suivante :

3.5.4 Traitement de la requête

L'indexation dans les systèmes de recherche d'informations classiques traite les termes de la requête extraits indépendamment de l'aspect sémantique de la requête. Or le système serait plus cohérent quand il présente à l'utilisateur les sources d'informations où les termes de la requête occurrent, et qui répondent à l'aspect sémantique de la requête. Ainsi que les documents couvrants le thème de cette requête. Il existe plusieurs moyens de récupération du thème d'une requête telle que la projection de la requête sur le réseau conceptuel d'une ontologie (Baziz, 2005) afin d'enrichir la requête avec

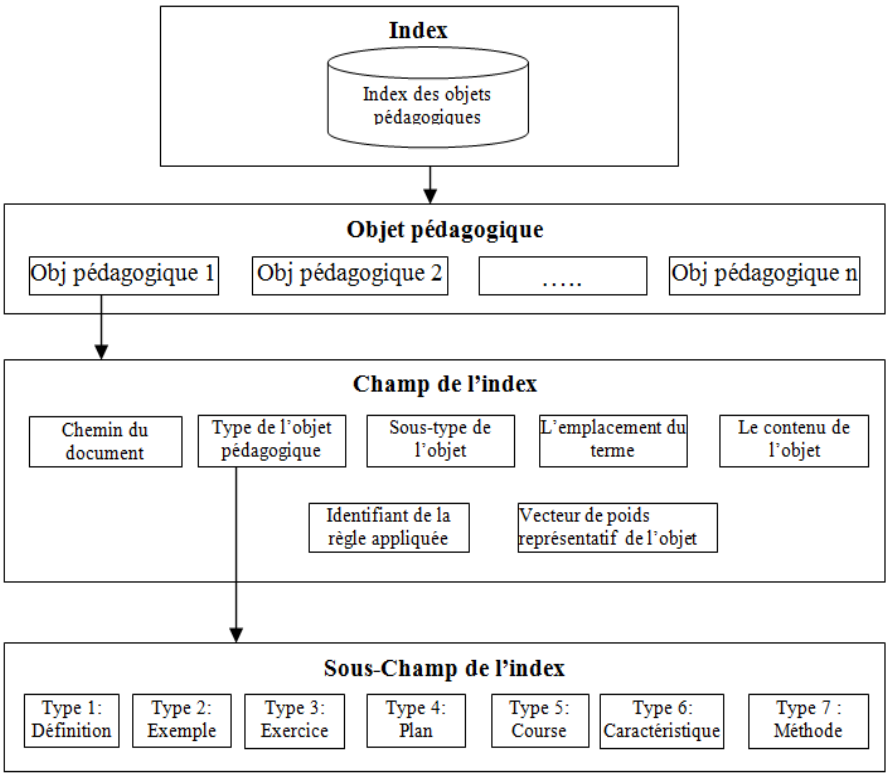


FIGURE 3.10: Organisation de l'index de SRIDOP

des termes apparentés pour élargir la recherche. Les résultats de ces méthodes d'indexation de la requête sont généralement excellents en rapidité de calcul, et acceptables en performances sémantiques (Memmi, 2000). Dans le cadre de notre système de recherche d'informations pédagogiques, nous proposons une indexation de la requête par une représentation vectorielle de ce dernier (Salton et al., 1975). C'est un vecteur composé des mesures de poids (fréquence) des termes index dans la requête. En plus de ce vecteur, le type de l'objet pédagogique choisit par l'utilisateur est un autre élément qui indexera l'aspect sémantique de la requête.

Le processus d'indexation de la requête suit les mêmes étapes que l'indexation des documents. Cependant, le processus d'indexation de la requête est plus simple car il s'agit simplement de détecter le choix du type de l'objet pédagogique effectué par l'utilisateur. Ce choix permet d'indexer la requête par le type d'objet pédagogique à rechercher.

Pour la vectorisation de la requête, le vecteur index d'une requête q est sous la forme suivante : $V_q = (p_q^1, p_q^2, p_q^3, ..., p_q^n)$

n est le nombre de termes index p_q^i est la mesure de poids (la fréquence) du terme i de la requête q. La requête sera indexée alors par des informations concernant son type pédagogique et son vecteur de poids.

Une fois la requête et les documents sont indexés, l'étape d'appariement peut débuter

pour sélectionner les objets pertinents par rapport à la requête utilisateur.

3.5.5 Appariement Document-requête

La recherche proprement dite s'effectue en calculant une mesure de similarité entre chaque document du corpus et la requête de l'utilisateur.
Dans notre cas, il s'agit de faire l'appariement entre les objets pédagogiques indexés et la requête utilisateur (Faiz et *al.*, 2012),(Smine et *al.*, 2013). Les objets ayant le même type que la requête sont extraits, ensuite l'appariement entre les objets et la requête est effectué en calculent la mesure de similarité entre les vecteurs de chaque objet et de la requête en utilisant la mesure Cosine de Salton (Salton et al., 1975).
Cette mesure est l'une des mesures de similarité les plus fréquemment utilisées grâce à son bon fonctionnement sur des corpus variés. Elle consiste à calculer les valeurs des cosinus des angles séparant les vecteurs des documents et le vecteur de la requête (selon le modèle vectoriel, les documents et la requête sont représentés dans le même espace). Par rapport à un simple produit scalaire, cette mesure présente l'avantage de normaliser les scores de chaque document en fonction de leur taille, elle-même pondérée par le poids des termes. La mesure Cosine est définie comme suit :

$$\cos(\overrightarrow{C}_{user}, \overrightarrow{D}) = \frac{\overrightarrow{C}_{user} \times \overrightarrow{D}}{\left|\overrightarrow{C}_{user}\right|\left|\overrightarrow{D}\right|}$$

Avec : $\overrightarrow{C}_{user}$ est le vecteur de la requête posée par l'utilisateur et $\overrightarrow{D}$ est le vecteur du document du corpus

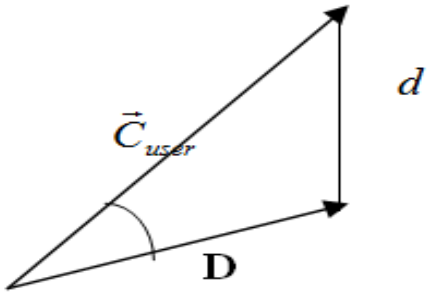


FIGURE 3.11: Le cosinus comme mesure de similarité ($\overrightarrow{C}_{user}$ et D sont respectivement les vecteurs représentant la requête et le document)

Il suffit donc d'examiner les composantes des vecteurs non nulles, à la fois pour la requête et le document. Dans le modèle vectoriel standard, cela revient à ne s'intéresser qu'aux mots partagés par la requête et le document, puisque les termes absents des documents (ou de la requête) ont une pondération nulle. Cela explique la mise en œuvre par fichiers inversés et la grande rapidité de cette étape d'appariement, notamment lorsque l'on manipule des requêtes de quelques mots.
Dans la figure X, nous avons repris l'exemple de l'étape de la représentation vectorielle pour montrer l'appariement entre les documents indexés et la requête selon le modèle vectoriel.

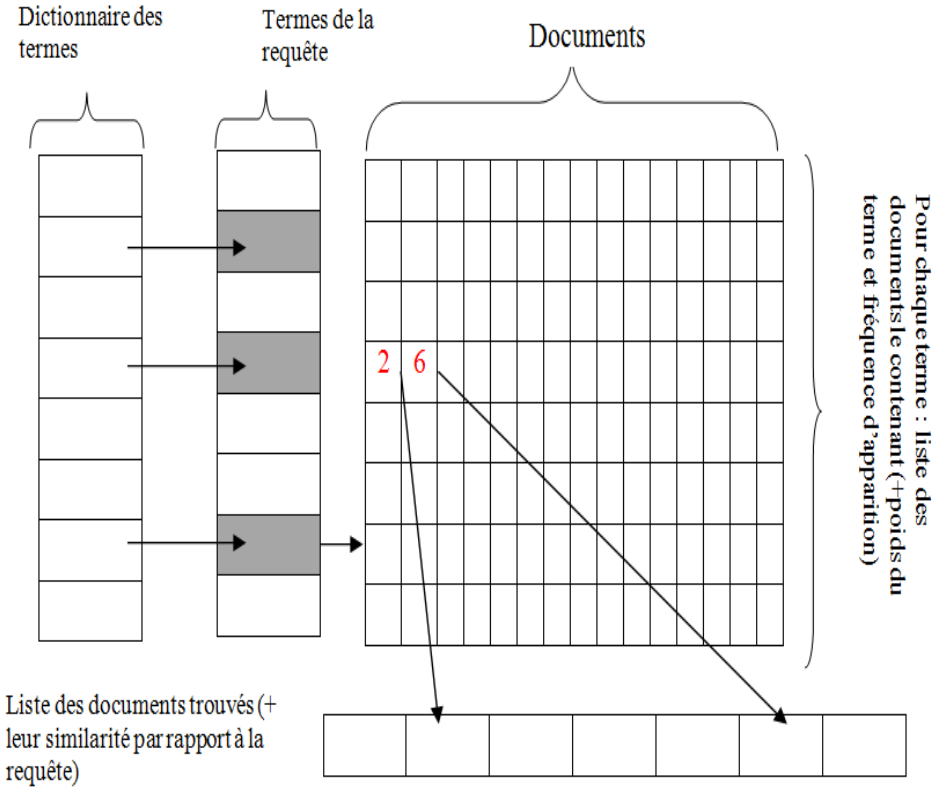


FIGURE 3.12: Exemple de fichier inversé (les documents 2 et 6 font partie des documents trouvés)

	L'informatisation	est	l'installation	d'un	système	de	traitement	automatique	l'information	la	mise	en	place	dans	un	service	n'utilisant	pas	l'informatique	au	préalable
Obj1	1	1	0	1	1	2	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Obj2	1	1	1	1	1	2	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0
Req	0	0	0	0	0	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	

Tableau 3.5: Les vecteurs des deux objets et celui de la requête

Nous présentons un exemple d'appariement entre les vecteurs des objets et de la requête en utilisant le modèle vectoriel. Soient les vecteurs des objets des exemples présentés précédemment, et nous ajoutons à ceux-ci le vecteur de la requête "traitement automatique de l'information".

Nous proposons dans ce qui suit un exemple de requête composé du terme "traitement automatique de l'information" et du type pédagogique ("Définition"). Pour extraire les réponses pertinentes, le moteur de recherche de SRIDOP procède comme suit :

1. Il extrait tous les objets pédagogiques trouvés dans l'index annotés avec le type pédagogique "Définition"
2. Il sélectionne ensuite les objets ayant une mesure de similarité avec la requête "traitement automatique de l'information"
3. Il affiche l'ensemble des informations relatives à chaque objet pédagogique trouvé

L'affichage des résultats se détaillé durant le chapitre suivant à travers les interfaces d'affichage.

3.5.6 La constitution de fiches pédagogiques

La constitution de fiches pédagogiques est un module qui suit la restitution des résultats aux utilisateurs. En effet, suite à l'affichage des résultats à l'utilisateur, nous lui donnons la possibilité d'afficher le corps des objets pédagogiques ainsi que leurs types dans un tableau dans une page sous format HTML. Nous avons nommé cette page "Fiche pédagogique" puisqu'elle rassemble les différents objets recherchés par l'utilisateur. Le

concept de "fiche pédagogique" facilite l'accessibilité aux différents résultats fournis à l'utilisateur.

3.6 Conclusion

Nous avons présenté dans ce chapitre un modèle d'annotation et d'extraction d'objets pédagogiques à partir de documents. Ce modèle est composé principalement d'une première partie d'indexation des objets pédagogiques suite à leur annotation sémantique, ainsi que à leur représentation vectorielle. Dans la deuxième partie, il s'agit d'indexer la requête introduite par l'utilisateur par le même procédé que celui d'indexation des documents. Dans une troisième partie, un appariement entre la requête et les documents est effectué en utilisant une fonction score. Les résultats pertinents représentent les objets pédagogiques répondant à la requête utilisateur, ainsi que les documents comportant ces objets. En plus du domaine pédagogique, notre modèle peut être appliqué à n'importe quel autre domaine, par exemple domaine de la biologie, des évènements, etc.
Nous avons implémenté le système SRIDOP pour valider notre modèle proposé. Ce système comporte un module d'indexation des documents (Annotation et Représentation vectorielle). Un autre module de gestion des règles d'Exploration Contextuelle et un troisième module de d'appariement entre les documents et la requête pour afficher les résultats pertinents par rapport à la requête utilisateur.
Dans le chapitre suivant, nous présentons les étapes d'implémentation de notre modèle ainsi que les outils utilisés pour l'implémenter. Nous illustrons aussi les résultats d'évaluation des différents modules composant notre système SRIDOP.

Chapitre 4

Mise en œuvre informatique et Evaluations

Sommaire

4.1 Introduction

L'objectif de ce travail est d'élaborer un outil de recherche d'informations pédagogiques à partir de documents. nous avons babtisé notre système : SRIDOP qui doit répondre à un certain nombres de requêtes formulées par l'utilisateur pour son besoin d'apprentissage, d'enseignement, de formation, etc. Nous avons implémenté notre approche à l'aide de techniques et de ressources informatiques récentes, connues et disponibles, afin de garantir la réutilisation de notre système ou de ces différents modules (Annotation, indexation, recherche d'objets pédagogiques, etc.), séparément.

Nous avons donc tenu à montrer l'adaptabilité de la méthode d'exploration contextuelle aux objectifs attendus, d'abord celui de l'annotation, ensuite de l'extraction des réponses pertinentes par rapport à la requête utilisateur. Notre but est de réaliser un système qui soit, le plus possible, complet et ergonomique. Nous avons essayé tout au long de la phase d'implémentation de choisir des méthodes informatiques efficaces (comme la méthode Orintée Objet), en termes de temps d'exécution et convivialité du système. Aussi nous avons choisie d'accorder beaucoup d'importance à l'interface de SRIDOP.

Durant ce chapitre, nous présentons, en premier lieu, les principales méthodes d'évaluation des systèmes de recherche d'informations. Nous exposons, ensuite, la mise en œuvre informatique de la formalisation étudiée au chapitre 3, en illustrant les résultats par des vues d'écrans obtenues lors de l'exécution du système SRIDOP. Nous donnons, en troisième lieu, un exemple complet de réalisation à l'aide des interfaces Homme machine de notre système SRIDOP. Enfin nous présentons les résultats d'évaluation des différents modules de notre système.

4.2 Evaluation des systèmes de RI

L'évaluation des systèmes de recherche d'informations peut être abordée selon deux angles : l'efficience et l'efficacité. L'efficience regroupe le temps et l'espace : plus le temps de réponse est court et plus l'espace occupé par le système est faible, meilleur est considéré le système. Ces critères ne concernent cependant que les systèmes qui assurent parfaitement une fonction précise, ce qui n'est pas le cas dans le domaine de la recherche d'information.

D'autres mesures de performances des SRI ont donc été introduites, dans le but d'évaluer l'efficacité des systèmes. Parmi elles, on peut citer la facilité d'utilisation du

système. Nous nous intéressons ici à celle qui nous semble la plus importante : la capacité d'un système à sélectionner des documents pertinents. Les mesures que nous présentons dans la suite de cette section rendent possible la comparaison des SRI entre eux. Cependant, pour que la comparaison soit valable, il faut que ces mesures soient effectuées dans les mêmes conditions. C'est de cette nécessité que sont nées de nombreuses campagnes d'évaluation, dont nous donnons un exemple dans la deuxième partie de cette section.
La performance des SRI est évaluée à partir de la pertinence des documents renvoyés. Cette notion de pertinence est ambiguë. En effet, on peut parler de pertinence objective, c'est-à-dire une pertinence calculée à partir des résultats du SRI, mais aussi de pertinence subjective : un document peut être jugé pertinent à une requête par un utilisateur et pas par un autre. De même, la pertinence d'un document dépend des connaissances de l'utilisateur sur le sujet, ce qui peut affecter la pertinence des documents examinés par la suite. C'est pour cette raison que des mesures d'évaluation orientées utilisateurs ont été introduites. Nous présentons ici les mesures d'évaluation de SRI les plus courantes, ainsi qu'un exemple de campagne d'évaluation utilisée par les centres de recherches pour comparer leurs différents systèmes.

4.2.1 Evaluation de la performance d'un système de recherche d'information

4.2.1.1 Rappel et Précision

D'une façon générale, tout SRI a deux objectifs principaux : retrouver tous les documents pertinents, et rejeter tous les documents non pertinents. Ces objectifs sont évalués par les mesures de rappel et de précision (Boughanem et al., 2008).
La précision mesure la proportion de documents pertinents relativement à l'ensemble des documents restitués par le système. Elle est exprimée par :

$$Précision = \frac{Documents\ pertinents\ sélectionnés}{Documents\ sélectionnés}$$

Le rappel mesure la proportion de documents pertinents restitués par le système relativement à l'ensemble des documents pertinents contenus dans la collection de documents. Il est exprimé par :

$$Rappel = \frac{Documents\ pertinents\ sélectionnés}{Documents\ pertinents}$$

L'avantage d'avoir les deux mesures de précision et de rappel, c'est que l'un est plus important que l'autre dans de nombreuses circonstances. Les utilisateurs voudraient avoir tous les résultats pertinents à la première page (haute précision), mais ils n'ont pas le

moindre intérêt à connaître tous les documents qui sont pertinents. En revanche, plusieurs chercheurs professionnels sont très préoccupés par essayer d'obtenir le rappel le plus élevé, en tolérant des résultats de précision assez faible pour l'obtenir. Néanmoins, les deux mesures ont tendance à fonctionner inversement : quand l'un augmente, l'autre diminue. En fait, on peut toujours obtenir un rappel de 1 (mais une précision très faible) en sélectionnant tous les documents pour toutes les requêtes et on peut obtenir une précision proche de 1 (mais un rappel très faible). Chaque système cherche généralement à avoir un équilibre entre ces deux valeurs, favorisant parfois l'une au détriment de l'autre, selon le but visé par le SRI.

4.2.1.2 F-mesure

Une mesure unique qui combine le rappel et la précision est le F-mesure, représentée par la formule suivante :

$$F = \frac{1}{\alpha\frac{1}{p}+(1-\alpha)\frac{1}{R}} = \frac{(\beta^2+1)PR}{\beta^2\ P+R} \quad avec\ \beta^2 = \frac{1-\alpha}{\alpha} \text{ où } \alpha \in [0,1]\ et\ \beta^2 \in [0,\infty]$$

La valeur de F-mesure égalise, par défaut, les mesures de précision et de rappel, en initialisant α à $\frac{1}{2}$ et β à 1. Dans ce cas la formule de F sera simplifiée :
$F = \frac{2PR}{P+R}$

Il existe d'autres mesures pour évaluer les systèmes de recherche d'information à savoir "Le bruit", "le silence", etc.

4.2.2 Collection de référence, un exemple : TREC

Les mesures d'évaluation des SRI permettent certes de les comparer, mais encore faut-il que les évaluations soient faites sur jeux de données identiques. De nombreux projets basés sur des corpus d'évaluation se multiplient depuis les années 70 (Boughanem et al., 2008).
Le projet le plus ambitieux est sans aucun doute le projet d'évaluation TREC (Text Retreival Conference) de la DARPA (Defense Advanced Research Project Agency). La campagne d'évaluation TREC, co-organisée par le NIST (National Institute of Standards and Technology) et la DARPA, a commencé en 1992. Elle a pour but d'encourager la recherche documentaire basée sur de grandes collections de test, tout en fournissant l'infrastructure nécessaire pour l'évaluation des méthodologies de recherche et de filtrage de l'information. Pour chaque session de TREC, un ensemble de documents et de requêtes (les topics) sont fournis. Les participants exploitent leurs propres systèmes de recherche

sur les données et renvoient au NIST une liste ordonnée de documents. NIST évalue ensuite les résultats comme suit. L'ensemble des documents pertinents pour chaque requête est obtenu en prenant les K documents les mieux classés des différents SRI participant à la campagne d'évaluation. Ces documents sont ensuite montrés à des juges qui décident finalement de la pertinence de chaque document. Les participants à TREC disposent de la liste des documents pertinents pour chaque requête, et peuvent ainsi évaluer les performances de leurs SRI respectifs.

D'autres campagnes d'évaluation ont vu le jour, nous citons en particulier la campagne INEX (INitaitive for the Evaluaton of the XML Retrieval) lancée en 2002 ; elle est destinée à construire des collections et métriques pour évaluer les travaux de RI sur XML.

4.3 Mise en œuvre informatique

4.3.1 Réalisation du système SRIDOP : Contraintes

Pour implémenter notre travail, nous avons été amenés à spécifier et à programmer les différents modules dont nous avions besoin pour les différentes tâches d'extractions. Les modules que nous avons développés sont plus précisément les modules de conversion, de segmentation des documents, le module d'annotation des objets pédagogiques, le module d'indexation des objets pédagogiques, le module de recherche des objets répondant à la requête utilisateur, le module de constitution de fiches pédagogiques et le module de gestion des règles d'Exploration Contextuelle.
La réalisation du système SRIDOP a été soumise à un ensemble de contraintes que nous nous sommes imposés afin de mener à bien ce travail. En définissant ces contraintes nous avons constitué en quelque sorte, le cahier des charges du système. Certaines de ces contraintes nous ont permis de fixer le cadre général de développement du système et d'autres nous ont permis d'établir le contexte interne de développement à atteindre. Nous avons cité ces différentes contraintes au cours de ce document (Chapitre 1), mais à titre indicatif nous pouvons les résumer ainsi :

4.3.1.1 Les contraintes de départ

- N'importe quel type d'utilisateur (apprenant, enseignant, etc.) d'un domaine donné doit pouvoir utiliser le système, du moment qu'il définit ses axes de recherches,

- Les documents à traiter doivent constituer un corpus homogène thématiquement en rapport avec la problématique de l'utilisateur,

- Les documents peuvent présenter une structure très variable : on peut avoir des textes ne présentant aucune structuration mais aussi des textes fortement structurés (titre, paragraphe ...), etc.

4.3.1.2 Les contraintes en sortie du système

- Le système doit répondre à la requête utilisateur en faisant une extraction des objets pédagogiques pertinents,

- Le système offre à l'utilisateur la possibilité d'une constitution d'une fiche pédagogique, item [-] Le système donne à l'utilisateur la possibilité de gérer les règles d'Exploration Contextuelle.

4.3.2 Les langages et outils utilisés

Nous avons choisit d'implémenter notre système SRIDOP en utilisant le langage JAVA. nous avons développé des classes dont l'extension ".java" et nous avons profité de l'aspect Orienté Objet du langage pour organiser notre projet en packages relatives aux différents modules de notre système.

Nous avons implémenté notre système SRIDOP sous l'environnement "Eclipse" qui permet de construire des applications Java. C'est un projet de la Fondation Eclipse visant à développer un environnement de développement intégré libre, extensible, universel et polyvalent. Son objectif est de produire et fournir des outils pour la réalisation de logiciels, englobant les activités de programmation (notamment au moyen d'un environnement de développement intégré) mais aussi de modélisation, de conception, etc.

Nous avons eu recours, dans la plus part des modules, à l'integration de quelques API (Application Programming Interface) pour accéder aux services offerts par l'environnement de développement. L'API est un ensemble de classes ayant pour objet de faciliter le travail d'un programmeur en lui fournissant les utils de base nécessaires à tout travail à l'aide d'un langage donné. Elle constitue une interface servant de fondement à un travail de programmation plus poussé. Par exemple, l'API "xml-apis" de documents XML. Nous avons aussi fait appel à deux API : Lucene et Digester.

Lucene est un API qui permet de créer un moteur de recherche libre pour d'indexer et de rechercher du texte. Apache Solr est basé sur la bibliothèque Lucene. Lucene offre de puissantes fonctionnalités qui justifient notre choix : il offre une indexation incrémentielle

aussi rapide que l'indexation des lots et la taille de l'index à peu près 20-30% de la taille du texte indexé. Il propose aussi une recherche par champ(par exemple, titre, auteur, contenu)et le tri par n'importe quel champ, ce qui nous a servi dans l'indexation des annotations des objets pédagogiques. Nous avons aussi profité des modèles de classement enfichables proposé par Lucene, y compris le modèle vectoriel, pour indexer les objets pédagogiques en utilisant le modèle vectoriel.

Pour l'indexation de nos documents XML avec Lucene, nous avons été amenés à faire appel à l'API Digester ayant pour but de créer des objets à partir de données contenues dans des fichiers XML. Nous avons fait appel à Digester pour créer des objets à la volée lorsqu'un certain enchaînement de balises est détecté (annotation d'un objet pédagogique)ou encore d'appeler des méthodes spécifiques sur ces objets en leur passant des paramètres issus du fichier XML.

Comme le suggère la structure d'un document XML, Digester fonctionne selon une logique arborescente. Le résultat de l'analyse d'un fichier est un élément racine à partir duquel il est possible d'accéder aux autres. Le raccordement des objets entre eux et à l'objet racine est à la charge de l'utilisateur. Il est possible de demander au Digester de créer des objets ; ceux-ci sont temporairement placés sur une pile (la pile d'objets) afin de pouvoir être utilisés mais à l'issue de l'analyse du fichier XML, seul l'élément racine est accessible. Il est possible donc de récupérer tous les objets à partir de l'objet racine, celui-ci peut tout simplement être une collection à laquelle peuvent être ajoutés les différents éléments au fur et à mesure de leur création.

4.3.3 Implémentation des modules

Pour implémenter notre travail, nous avons été amenés à spécifier et à programmer les différents modules dont nous avons besoin pour les différentes tâches de recherche d'informations pédagogiques (cf. Fig.4.1). Les modules que nous avons développés sont les suivants :

- Conversion en texte
- Segmentation
- Annotation des objets pédagogiques
- Indexation des objets pédagogiques
- Recherche des objets pédagogiques répondant à la requête utilisateur
- Constitution de fiches pédagogiques

- Gestion des règles d'exploration contextuelle

La figure suivante (cf. Fig.4.1) illustre les différents modules implémentés et leurs relations les uns avec les autres : les rectangles à coins arrondis sont les différents modules et les formes cylindriques sont soit des ressources linguistiques et ls documents, soit les résultats des modules.

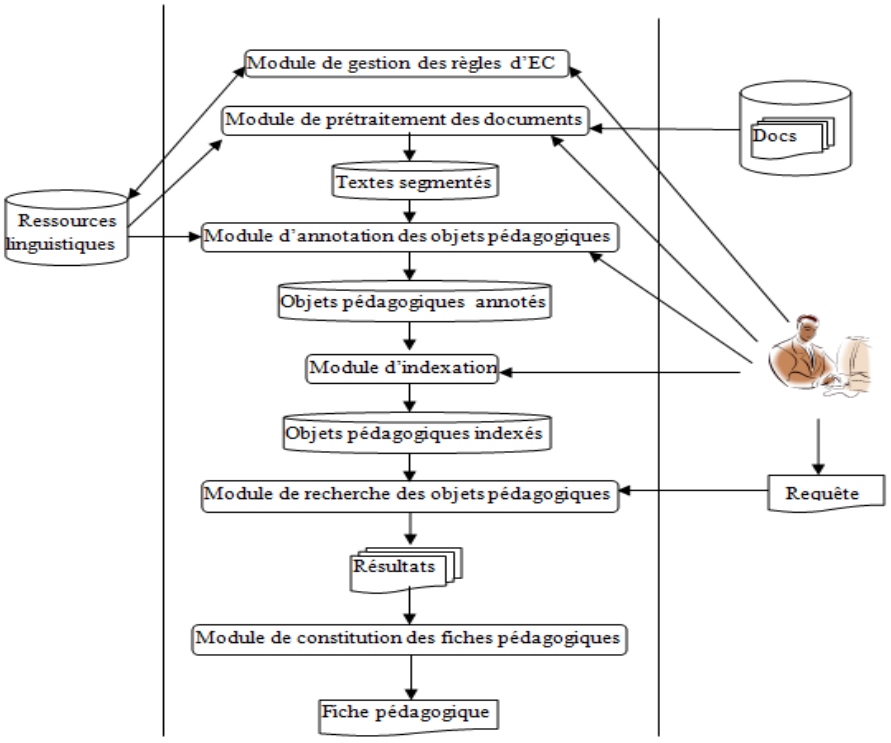


FIGURE 4.1: Architecture du système SRIDOP

Dans le cadre de notre système, tous ces modules sont enchaînables mais non dépendants. En fait, l'exécution d'un module se fait sur les résultats du module qui lui précède. Par exemple, le module d'annotation des objets pédagogiques ne peut s'exécuter que sur des documents segmentés. Sauf, le module de gestion des règles d'EC qui est indépendants des autres modules. L'indépendance des modules a fait que l'exécution d'un seul module peut se faire sans la nécessité d'exécuter les autres modules.

Tous les modules de notre système sont aussi importants les uns que les autres, sauf que les principaux modules sont : l'annotation des objets pédagogiques et la recherche des objets pertinents répondant à une requête utilisateur. Les autres modules (Gestion des règles d'EC, Segmentation, Constitution de fiches pédagogiques, etc.) sont des modules

qui synchronisent ou complètent les fonctionnalités offertes par notre système comme le module de constitution de fiches pédagogiques qui offre davantage d'autres services à l'utilisateur.

La fenêtre principale de notre système SRIDOP (cf. Fig.4.2) est la première fenêtre qui rassemble toutes les fonctionnalités dans son menu.

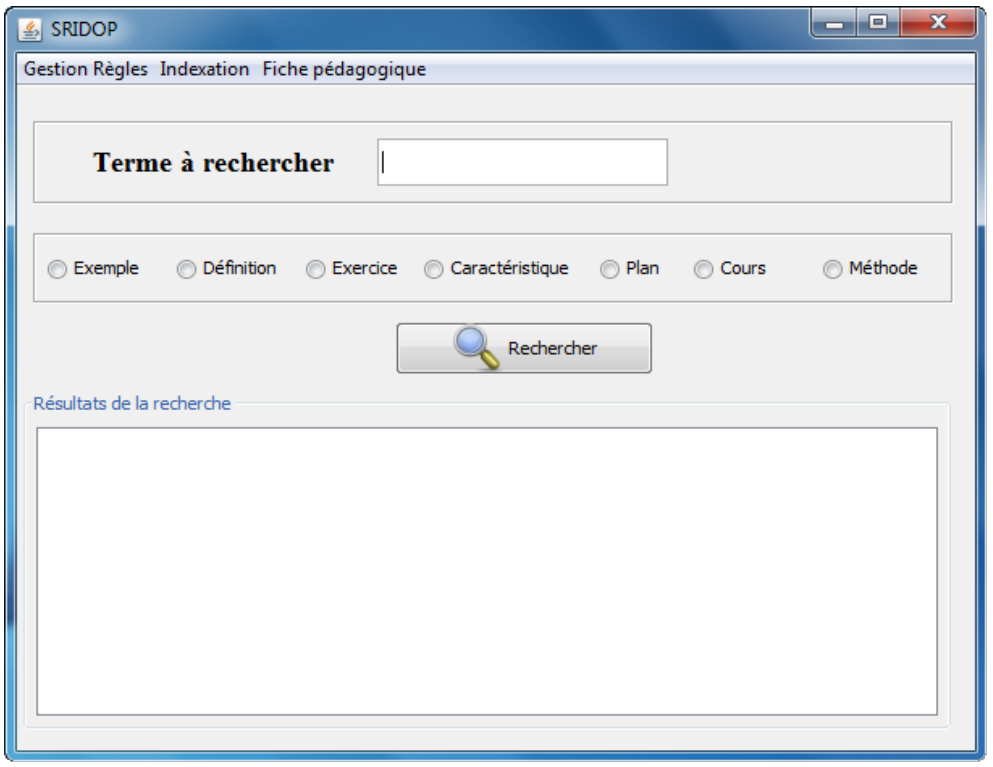


FIGURE 4.2: Fenêtre principale du système SRIDOP

- Le menu "Gestion Règles" est constitué des commandes "ajouter règle", "modifier règle" et "consulter règle".
- Le menu "Indexation" comporte les fonctionnalités "Conversion en fichiers TXT", "Segmentation", "Annotation" et "Indexation avec Lucene".
- Le menu "Fiche pédagogique" est constitué de la fonctionnalité "Constitution de fiche pédagogique".

La principale fonctionnalité de notre système consiste en la recherche des objets pédagogiques répondant à la requête utilisateur. L'utilisateur doit saisir le terme à rechercher dans le champ destiné à la saisie de ce dernier. Ce terme doit correspondre

au contenu de l'objet pédagogique retourné. Il doit ensuite choisir le type des objets pédagogiques recherchés. Les résultats sont affichés dans le champ en bas destiné à l'affichage des résultats.

Nous présentons dans la figure qui suit un diagramme d'activité qui décrit le processus d'exécution des différentes tâches par l'utilisateur.

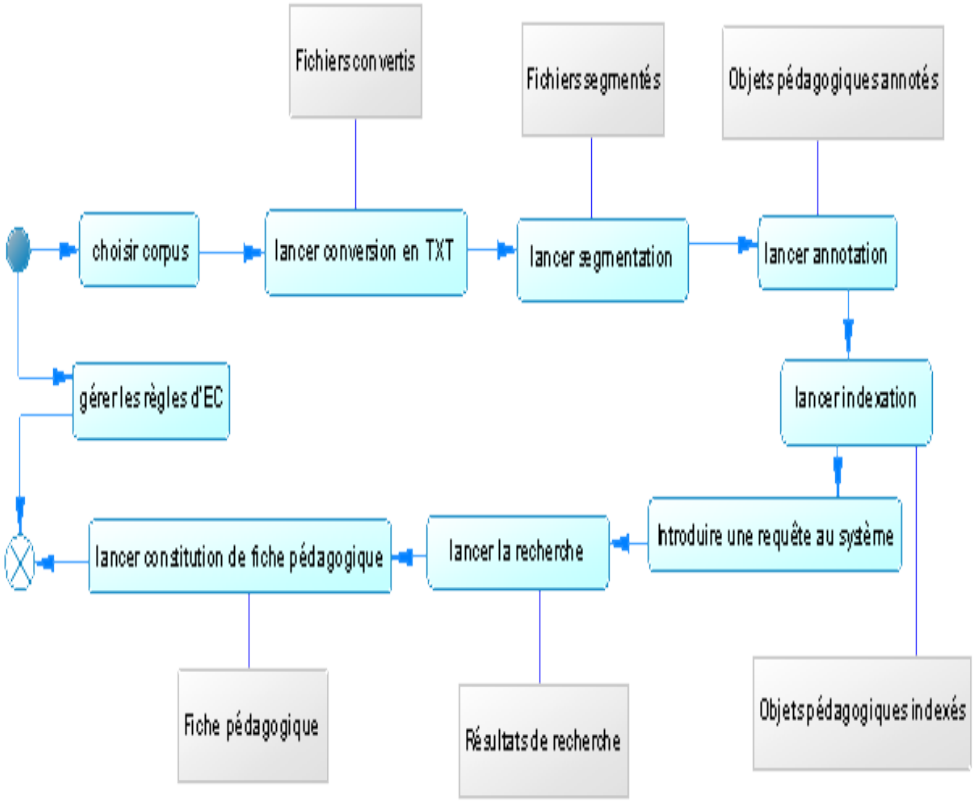


FIGURE 4.3: Fenêtre pour le lancement du module d'indexation

Le diagramme d'activité (cf. Fig.4.16) présente le chemin des activités effectué par un utilisateur. Les rectangles à coins arrondis représentent les activités. Les rectangles à coins droits sont les objets résultats de l'activité effectuée. L'activité "gérer les règles d'EC" est une activité indépendante du chemin d'annotation et d'indexation des objets pédagogiques.

Nous avons développé cinq paquetages (packages) qui renferment plus de 15 classes principales dans lesquelles sont instanciés les différents objets, nécessaires au fonctionnement du système SRIDOP :

- Le *package* “Gestion des règles d’EC44 qui prend en charge le module permettant de gérer les règles d’exploration contextuelle : ajouter règle, consulter règle et modifier règle.
- Le *package* “Conversion en fichiers TXT” qui prend en charge le module de conversion des fichiers de différents formats en fichiers textes.
- Le *package* “Segmentation” qui s’intéresse au module de segmentation des textes en sections, paragraphes et phrases.
- Le *package* “Annotation” pour l’annotation sémantique des objets pédagogiques.
- Le *package* “Recherche et Extraction des objets” qui prend en charge trois modules, à savoir : (a) Le module d’indexation des objets pédagogiques, (b) Le module d’extraction des objets pédagogiques répondant à une requête utilisateur, (c) Le module de constitution d’une fiche pédagogique rassemblant les objets pédagogiques.

La figure suivante (cf. Fig.4.3) présente une vue globale des interactions entre les différents paquetages du système SRIDOP. Nous allons revenir, dans les sections suivantes, sur chacun des packages développées lors de l’implémentation du système SRIDOP.

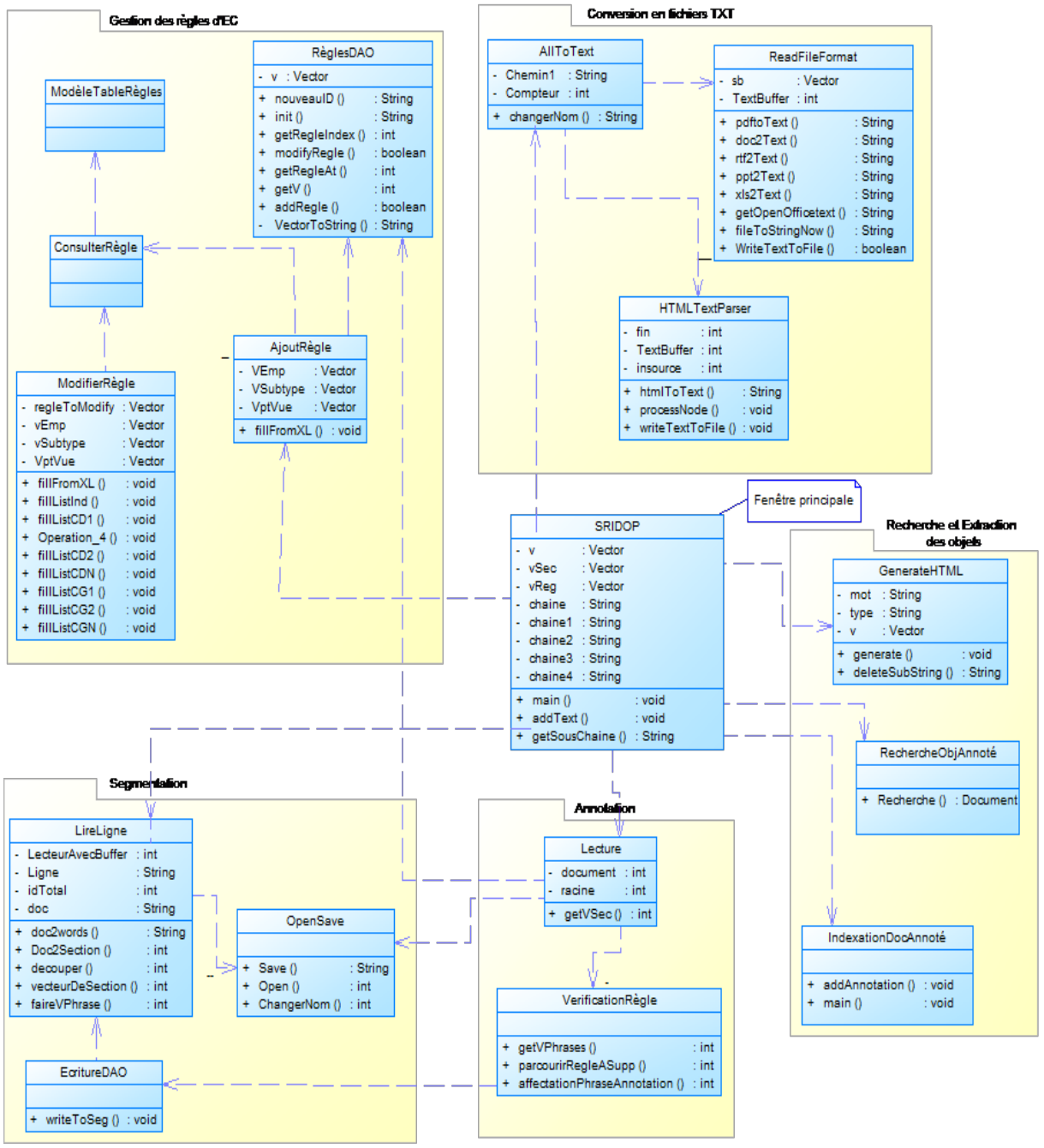


FIGURE 4.4: Diagramme de paquetages du système SRIDOP

Les différents modules implémentés seront détaillés dans ce qui suit.

4.3.3.1 Le module de Gestion des règles d'Exploration Contextuelle

Le module de "Gestion des règles d'Exploration Contextuelle" permet de gérer les règles à travers des interfaces. Les règles sont stockées dans un fichier Excel nommé

"Règles.xls". Nous présentons ci-dessous le diagramme de composants de ce module :

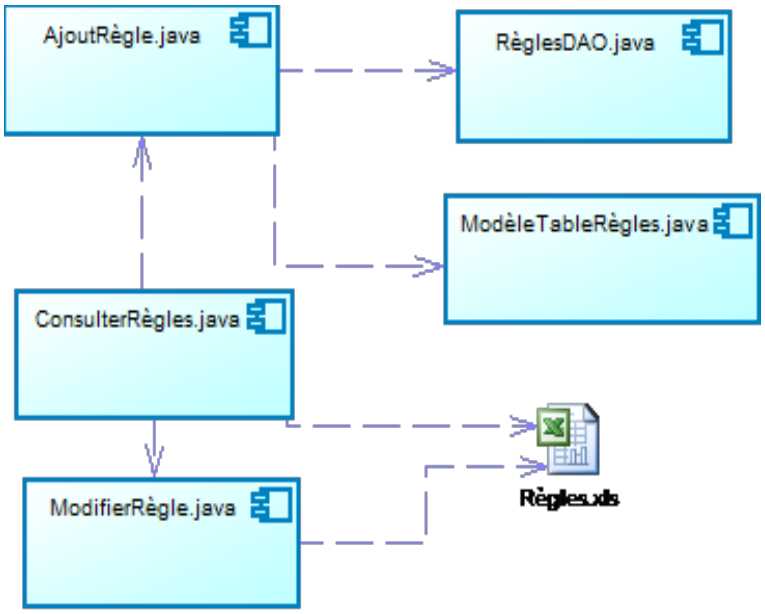


FIGURE 4.5: Diagramme de composants du module "Gestion des règles d'EC"

Ce module se base sur cinq classes :

- La classe "AjoutRègle.java" : elle permet d'ajouter une règle au fichier des règles
- La classe "ConsulterRègles.java" : elle permet de consulter une règle
- La classe "ModifierRègle.java" : elle permet de modifier une règle
- La classe "RèglesDAO.java" : cette classe permet de représenter les règles du fichier Excel sous la forme de vecteurs
- La classe "ModèleTableRègles.java" : cette classe permet la mise en forme des données saisies par l'utilisateur dans l'interface SRIDOP pour les transférer dans le fichier Excel.

4.3.3.2 Le module de conversion des fichiers sourcesen fichiers textes

Ce module prend en entrée un fichier source et produit en sortie un fichier texte contenant du texte brut. Ce module se base sur trois classes principales :

- La classe "AllToText.java" : Cette classe intervient pour déclencher le processus de conversion au début et pour changer le nom du fichier converti à la fin

- La classe “ReadFileFormat.java” convertie les fichiers de différents formats (DOCX, PDF, PPT, RTF, etc.)
- La classe “HTMLTextParser.java” permet de parcourir les fichiers HTML et de les convertir en fichiers texte.

La figure suivante (cf.Fig.4.5) présente le diagramme de composants du module de conversion

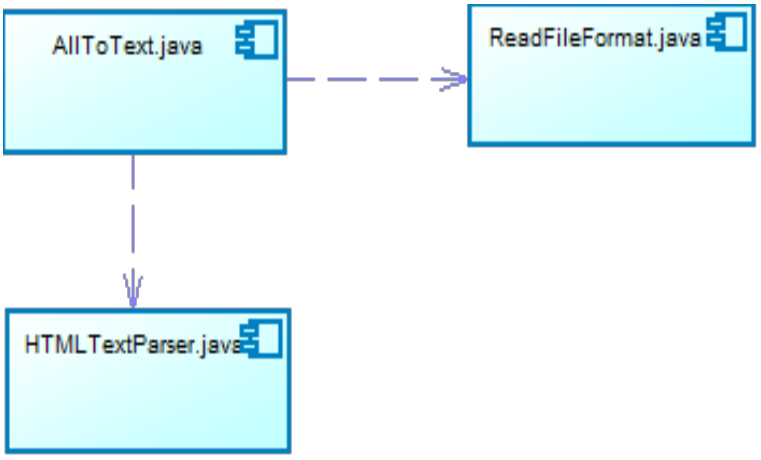


FIGURE 4.6: Diagramme de composants du module “Conversion”

4.3.3.3 Le module de segmentation

Ce module prend en entrée le texte brut qui est produit par le module de conversion et applique un algorithme basé sur le principe d’exploration contextuelle pour segmenter le texte en sections, paragraphes, phrases, etc. Le fichier, produit en sortie, contient le texte initial segmenté en format XML. Ce module se base sur les classes suivantes :

- La classe “LireLigne.java” permet de lire le fichier texte et d’appliquer les règles de segmentation qui sont représentés sous forme de vecteurs.
- La classe “OpenSave.java” permet de choisir le chemin des fichiers à segmenter et le chemin de l’enregistrement des fichiers segmentés
- La classe “EcritureDAO.java” permet d’écrire dans le nouveau fichier segmenté

La figure suivante (cf.Fig.4.6) présente le diagramme de composants du module de segmentation

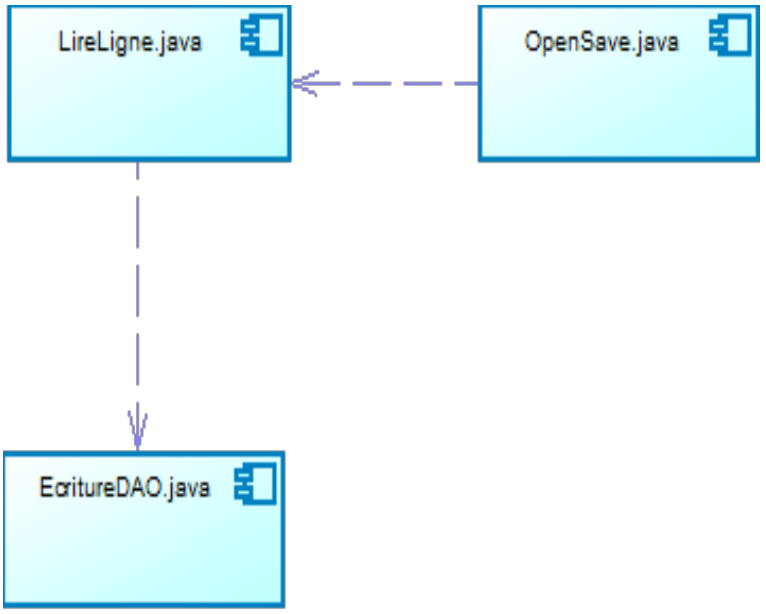


FIGURE 4.7: Diagramme de composants du module "Segmentation"

4.3.3.4 Le module d'annotation sémantique et automatique des objets pédagogiques

Le module d'annotation prend en entrée des fichiers segmentés en format XML et produit des textes annotés contenant des balises d'annotation à chaque fois un objet pédagogique est repéré dans le texte. Ce module se base sur les classes suivantes :

- La classe "Lecture.java" permet de lire le contenu du fichier segmenté en vue de son annotation
- La classe "OpenSave.java" et la classe "RèglesDAO.java" ont le même rôle que celui énoncé auparavant
- La classe "VerificationRègle.java" : c'est la classe principale qui permet d'insérer une annotation dans le fichier segmenté à chaque fois un objet est repéré selon la présence des éléments d'une règle.

Nous présentons ci-dessous le diagramme de composants relatif au module d'annotation (cf.Fig.4.7) :

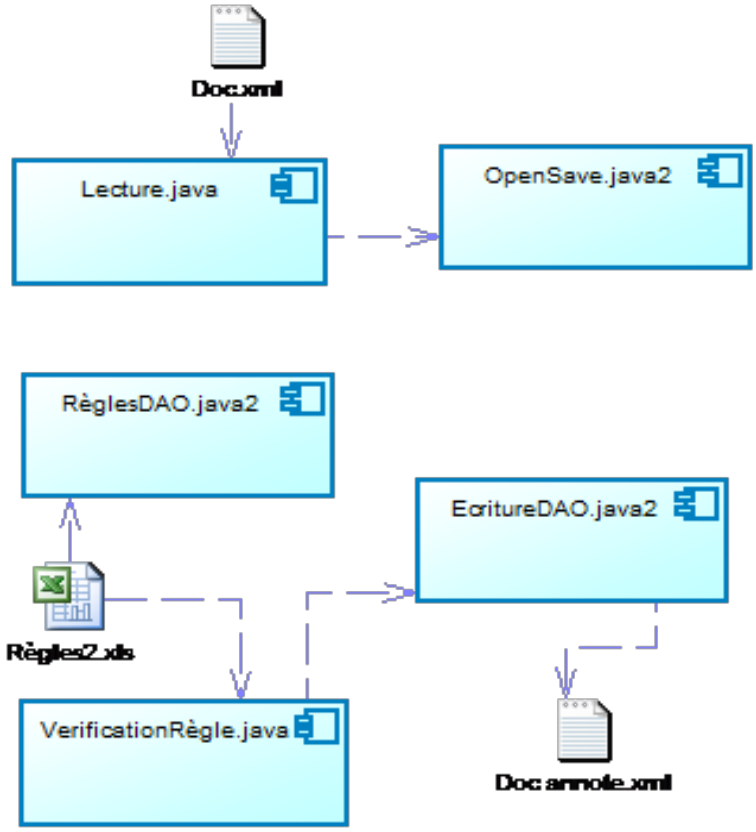


FIGURE 4.8: Diagramme de composants du module "Annotation des objets pédagogiques"

4.3.3.5 Le module de recherche et d'extraction des objets pédagogiques

Ce module est composé de trois parties :

- **Indexation des objets pédagogiques**

  Le module d'indexation des documents pédagogiques se base sur la classe "IndexationDocAnnoté". Dans cette classe d'indexation, nous avons utilisé les deux bibliothèques *Lucene* et *Digester* pour indexer les fichiers XML annotés. En effet, *Lucene* est un moteur d'indexation principalement des fichiers texte. Puisque nos fichiers sont de type XML, nous utilisons, en plus de *Lucene*, la bibliothèque *Digester* pour les indexer. En effet, *Digester* est une bibliothèque qui permet l'indexation des fichiers XML à l'aide de *Lucene*.

  La classe "IndexationDocAnnoté" contient la méthode "addAnnotation" qui permet d'ajouter un objet annoté à l'index à chaque une balise d'annotation est repérée dans le fichier annoté. Le fichier index est organisé en termes d'objets pédagogiques et non pas en termes de documents.

- **Recherche des objets pédagogiques répondant à la requête utilisateur**

  Ce module est le module noyau de notre application SRIDOP. Il offre la fonctionnalité principale de notre système à savoir recherche et extraction des objets répondant à la requête utilisateur. La classe principale de ce système est nommée "**RechercheObjAnnoté**". Elle permet de faire un appariement entre la requête utilisateur et les objets stockés dans l'index. Ensuite, elle permet une extraction des objets pertinents. Les objets pertinents sont ceux qui sont de même type que celui choisi par l'utilisateur et possédant un vecteur de fréquence des termes ayant une mesure de similarité avec le vecteur de la requête.

- **Constitution de fiches pédagogiques**

  Ce dernier module permet de générer une fiche pédagogique contenant les résultats répondant à la requête utilisateur. Cette fiche est sous la forme d'un fichier HTML contenant un tableau qui résume le contenu des objets pédagogiques pertinents, leurs types et leurs sous-types. La classe qui permet cette constitution est nommée "GenerateHTML".

4.3.4 Interfaces Homme Machine

Dans cette section, nous présentons les différentes interfaces des différents modules de notre système SRIDOP.

– **Le Module de gestion des règles**

Le module de gestion des règles d'EC est composé de trois sous-modules à savoir : "Ajouter Règle", "Modifier Règle" et "Consulter Règle". Dans ce qui suit, nous présentons l'interface relative à l'ajout d'une règle. Suite à la sélection du Menu Gestion Règles et de l'option "Ajouter Règle", l'interface suivante (cf. Fig.4.8) s'affiche pour permettre à l'utilisateur d'ajouter une règle d'exploration contextuelle

Suite à la saisie des différents constituants de la règle, l'utilisateur valide les données saisies. La classe "AjoutRègle" est instanciée pour ajouter les constituants dans leurs champs relatifs dans le fichier Excel où sont stockées les règles.

Pour le reste des modules de notre système à savoir : "Conversion en texte", "Segmentation", "Annotation", "Indexation avec Lucene", le système propose à l'utilisateur de choisir le chemin du corpus à traiter.

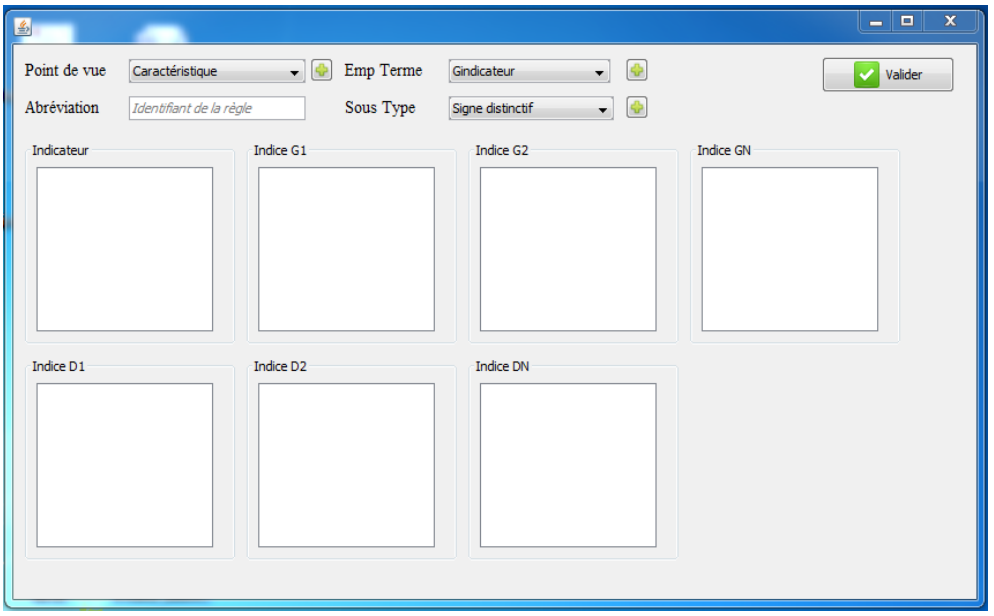


FIGURE 4.9: Interface d'ajout d'une règle d'EC

Pour une meilleure présentation des modules, nous avons choisit de poursuivre avec un extrait d'un document pédagogique. L'exemple est donné ci-dessous (cf. Fig.4.9). C'est un extrait du document pédagogique n°4 de l'annexe A.

– **Le module de conversion des fichiers**

Le résultat de la conversion de l'extrait textuel présenté précédemment est le suivant (cf. Fig.4.10) :

– **Le module de segmentation**

Le résultat de segmentation de l'extrait textuel est le suivant (cf. Fig.4.11) :

– **Le module d'annotation**

Nous présentons ci-dessous un extrait du résultat d'annotation de l'extrait textuel précédent (cf. Fig.4.12) :

– **Le module d'indexation des objets pédagogiques**

La procédure d'indexation s'exécute comme suit : Quand l'utilisateur choisit la commande "Indexation avec Lucene" (cf. Fig.4.13), une deuxième fenêtre s'affiche pour permettre à l'utilisateur de choisir le corpus à indexer.

1 – La portée des contraintes.

Dans une application qui utilise SQL, on trouve les éléments suivants : des bases de données relationnelles dans lesquelles se trouvent des tables et des vues. Tables et vues sont dotées de colonnes et les données sont écrites lignes par lignes. Finalement l'élément le plus petit de cet ensemble est la donnée. Les contraintes se trouvent à chacun des niveaux de cet édifice.
Les contraintes dites "de domaine" concernent les valeurs que revêtent les données des colonnes. Les contraintes de tables peuvent porter sur une colonne, sur une ligne ou sur une table, mais valident la cohérence de la ligne. Enfin les assertions peuvent porter sur plusieurs tables, voire toutes les tables de la base et assurent une cohérence transverse.

1.1 - Portée des contraintes de table

Les contraintes de table sont les plus connues. La plus classique ne concerne qu'une colonne à la fois. C'est la contrainte d'**obligation de valeur** (NOT NULL) qui exige, pour la colonne qui en est pourvue, qu'à toute ligne de la table une valeur soit exprimée.
On trouve ensuite la contrainte de **clef primaire** (PRIMARY KEY) qui assure l'unicité de la référence à une ligne d'une table. C'est le moyen par lequel on repère une ligne et une seule dans la table. Toutes les colonnes concourant à la clef se doivent d'être valuées (NOT NULL).
La contrainte d'**unicité** (UNIQUE) permet de s'assurer qu'une autre clef pourrait remplacer la clef primaire. Mais à la différence de la clef primaire, la contrainte d'unicité n'oblige pas à ce que les données participant à la formation de la contrainte soient valuées. Il peut même y avoir plusieurs lignes de la table dont les données formant la contrainte d'unicité sont vides de toutes valeurs. En d'autres termes, la contrainte d'unicité exige que toute donnée *valuée* soit distincte... Par essence une donnée non valuée ne peut jamais être comparée à une autre données non valuée, pas même à elle même. C'est à dire que le prédicat " NULL = NULL " ne sera ni vrai ni faux mais tout simplement inévaluable !
La contrainte la plus draconienne est la contrainte de **validation** (CHECK). Elle permet de restreindre les valeurs de la ou les colonnes qui la composent afin de respecter des règles

FIGURE 4.10: Extrait d'un document

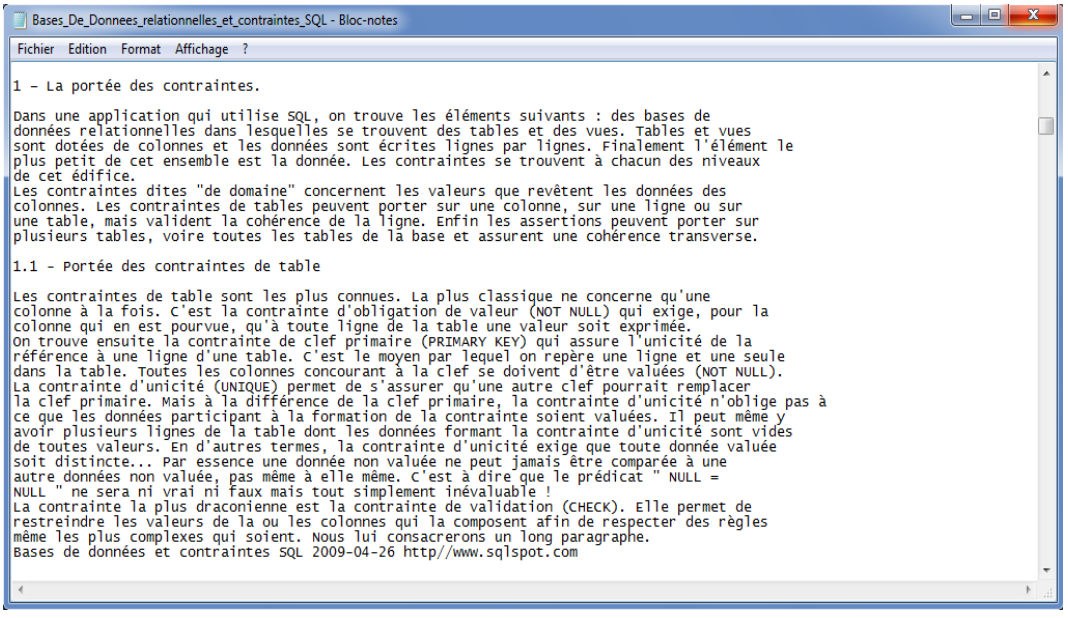


FIGURE 4.11: Résultat de conversion en texte brut de l'extrait

FIGURE 4.12: Résultat de segmentation de l'extrait

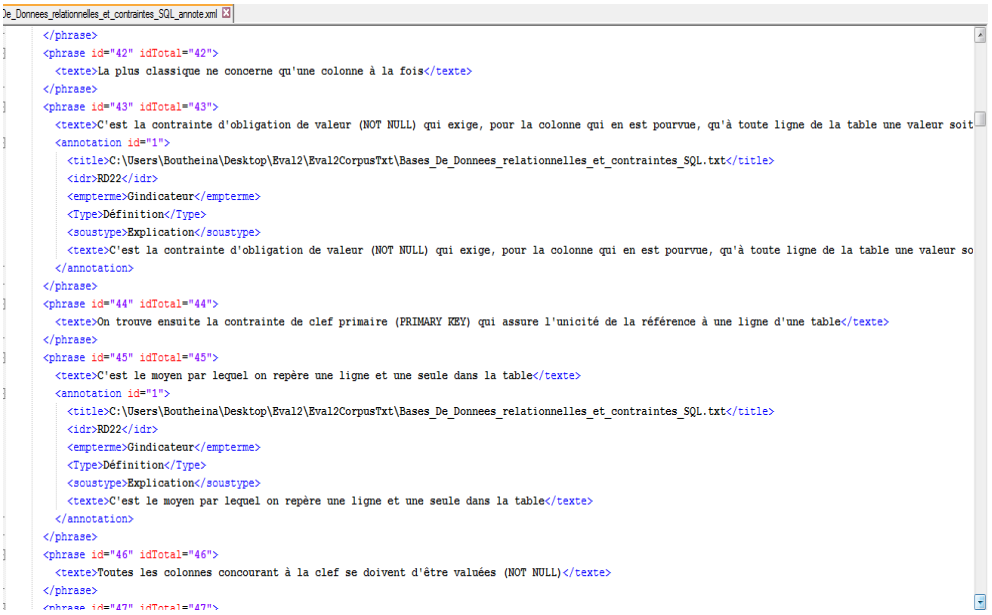


FIGURE 4.13: Résultat d'annotation de l'extrait

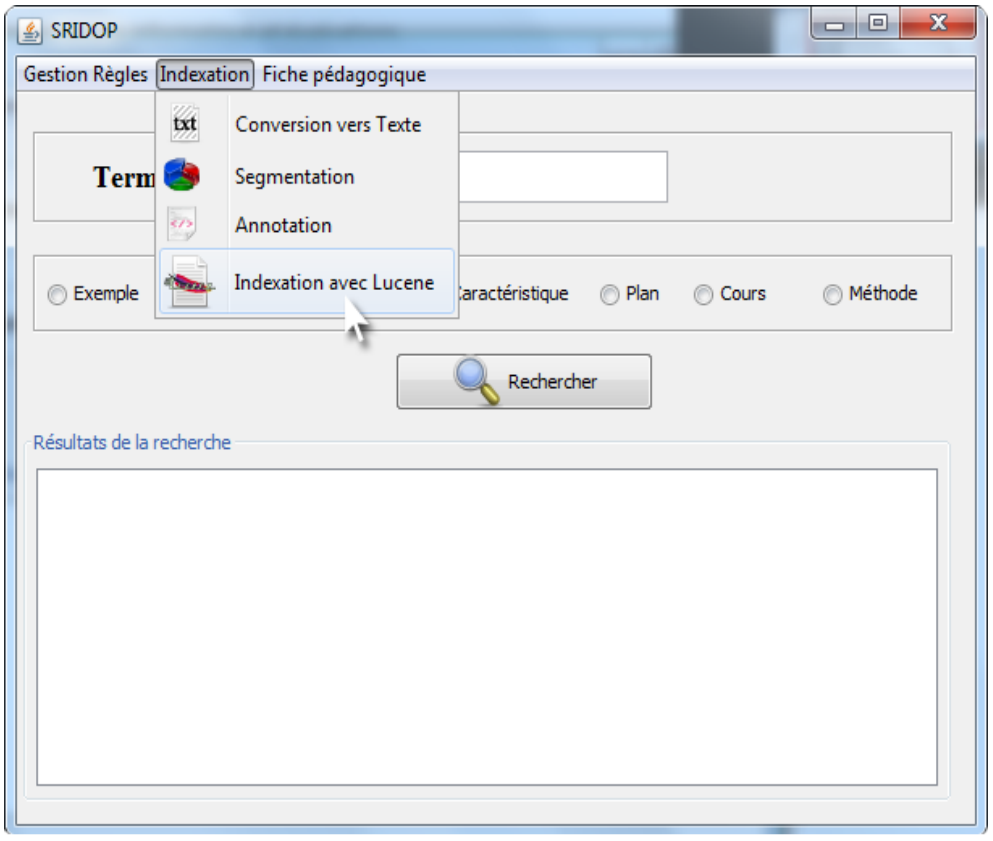


FIGURE 4.14: Fenêtre pour le lancement du module d'indexation

L'indexation aboutit à une création de fichiers sous le chemin "C :/Fileindexer". Ces fichiers représentent l'index des documents sélectionnés. Une fois l'index construit, le système recherche les résultats répondant à la requête utilisateur.

– **Le module d'extraction des objets pédagogiques répondant à une requête utilisateur**

L'utilisateur pose sa requête à travers l'interface principale du système SRIDOP. Un exemple ci-essous d'une requête posée par l'utilisateur représentée dans l'interface suivante :

– **Le module de constitution des fiches pédagogiques**

Suite à la sélection de la commande "Constitution de fiche pédagogiques", un fichier HTML est créé dans lequel est inséré un tableau contenant le contenu des objets pédagogiques retournés comme résultats à la requête de l'utilisateur, leurs types et leurs sous-types.

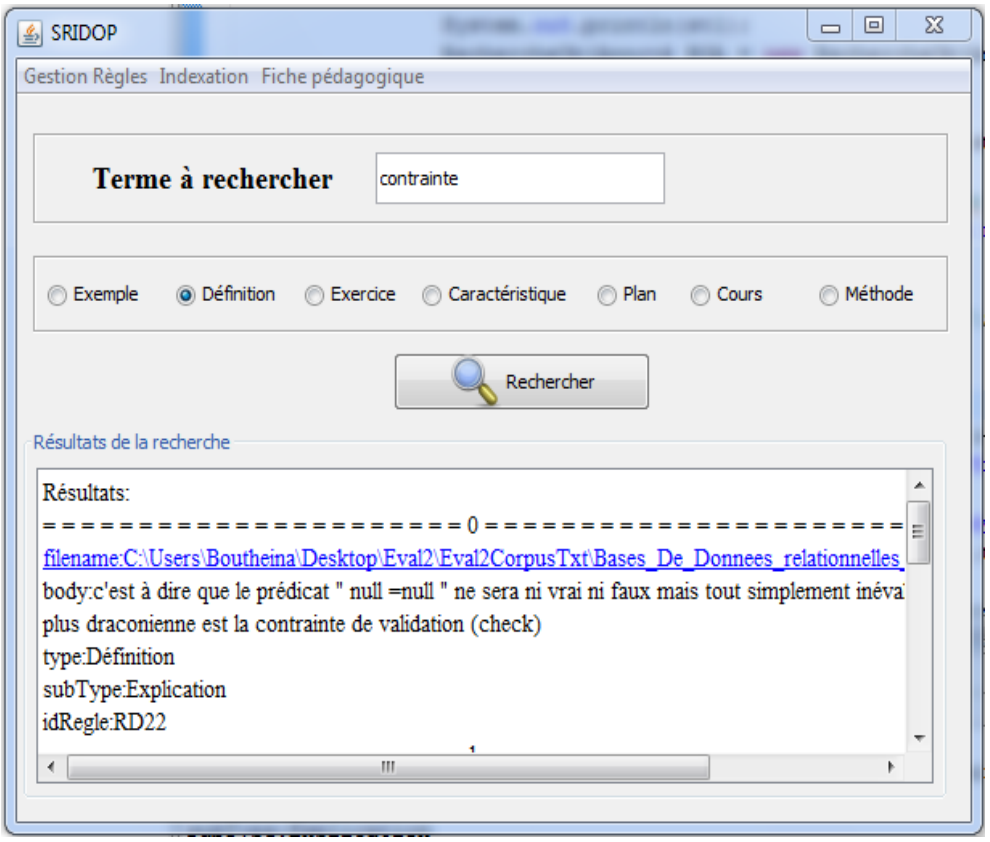


FIGURE 4.15: Fenêtre pour le lancement du module d'indexation

4.4 Evaluation de notre système SRIDOP

L'évaluation des systèmes a pris une place très importante dans le domaine des réalisations informatiques. Aujourd'hui, il n'est plus possible de proposer une méthode qui n'ait pas fait l'objet d'une quelconque évaluation. La difficulté est que chaque système a des spécificités qui le rendent difficilement comparable à d'autres systèmes, mêmes semblables. Ainsi, de nombreux systèmes visent des objectifs particuliers et nécessitent la mise en place de méthodes d'évaluation spécifiques. Par contre, d'autres systèmes, qui ont des objectifs communs, peuvent intégrer des programmes d'évaluations. Nous avons opté, dans notre travail, pour une évaluation qualitative et une autre quantitative. Pour l'évaluation qualitative, le système est évalué sur ses qualités en termes de son utilité, sa facilité d'utilisation, son ergonomie, etc. Dans l'évaluation quantitative, le système est soumis à une évaluation des résultats retournés en termes de pertinence. Ces deux types d'évaluations seront détaillés dans ce qui suit :

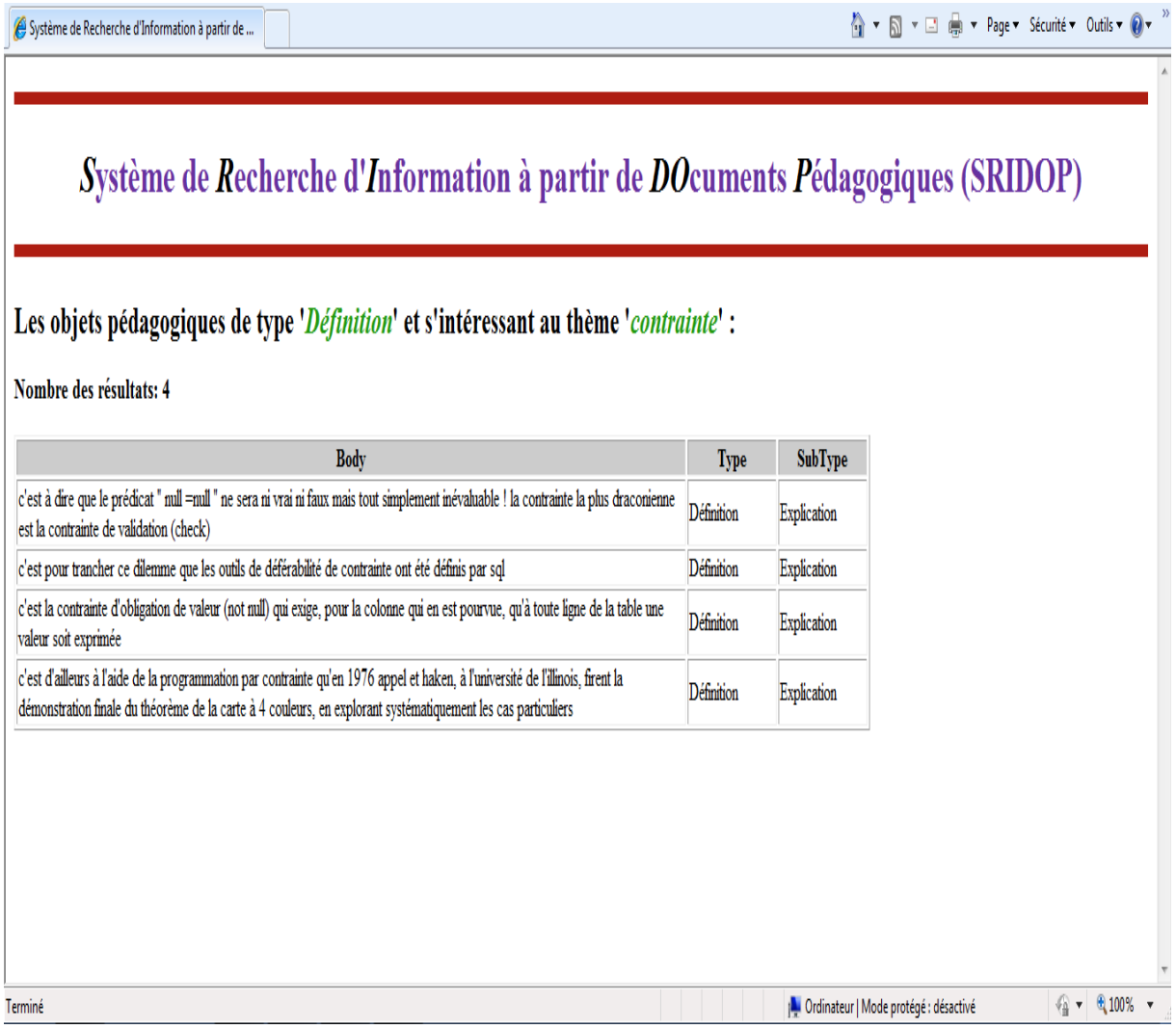


FIGURE 4.16: Exemple d'une fiche pédagogique

4.4.1 Evaluation qualitative

L'objectif de cette évaluation est de vérifier l'utilité de notre système, ainsi que l'intérêt de l'application de ses différents modules dans un scénario bien déterminé. Pour ceci, nous avons présenté le système à un enseignant en biologie (évaluateur) qui a un problème pédagogique concernant la construction de son cours. D'après ses révélations, il a appliqué le système sur son corpus situé sur son ordinateur. Il a tout d'abord effectué le prétraitement de son corpus, ensuite il a lancé l'annotation et l'indexation de son corpus. Finalement, il a exécuté des requêtes nécessaires à la constitution de cours (Plan du cours, Exemples sur une notion, Définition d'un concept, etc.) et a lancé, suite à chaque requête, la constitution d'une fiche pédagogique pour synthétiser les résultats. Dans certains cas, il a eu besoin de retourner au fichier principal pour voir le contexte dans lequel est situé l'objet

pédagogique.

Suite à l'utilisation de notre système, l'évaluateur a confirmé que le système est utile et présente une aide considérable, du moins dans le scénario dans lequel est appliqué. Il a ajouté qu'il est facile dans son utilisation, ergonomique, rapide dans l'exécution des tâches, etc. Nous présentons dans la figure qui suit un diagramme d'activité qui décrit le processus d'exécution des différentes tâches par l'utilisateur, ergonomique, rapide dans l'exécution des tâches, etc.

4.4.2 Evaluation quantitative

4.4.2.1 Les difficultés de l'évaluation

Une évaluation doit s'appliquer sur un nouveau corpus de documents, c'est-à-dire sur un corpus qui n'a pas été utilisé pour mettre au point la méthode évaluée. Ce corpus doit permettre l'obtention d'un minimum de résultats permettant une évaluation.

L'évaluation consiste à comparer les résultats obtenus selon la méthode testée sur le corpus d'évaluation avec des résultats choisis par un évaluateur (utilisateur potentiel, spécialiste du domaine ou un linguiste en fonction de la méthode). Ainsi, plus le corpus est volumineux et la méthode complexe, plus le temps nécessaire à l'évaluateur est important. Si les applications de la méthode évaluée sont restreintes, le nombre d'évaluations sera réduit. Par contre, si la méthode vise une application large, il se peut qu'un grand nombre d'évaluations soit nécessaire : méthode adaptée à toutes les langues, à tous les styles de rédaction, à tous les domaines, etc. Or, il n'est pas possible d'évaluer une méthode pour toutes les langues, ou pour les domaines (il faudrait disposer de corpus correspondants à ces critères). Par ailleurs, il est fréquent d'adapter le protocole d'évaluation à la méthode testée.

Enfin, il n'est pas évident de disposer de personnes objectives qui sélectionnent un corpus, proposent un protocole d'évaluation, et qui analysent les textes. En général, le choix du corpus et la description du protocole d'évaluation reviennent à la personne qui a développé la méthode testée, car c'est elle qui a le plus de temps à consacrer à la validation de son approche. Cela remet un peu en cause l'objectivité de l'évaluation.

4.4.2.2 La méthode d'évaluation retenue pour notre système SRIDOP

Une tentative d'évaluation a été effectuée lors du développement du système SRIDOP. Nous donnons ici les principaux résultats ainsi qu'une analyse critique de l'évaluation chiffrée qui a été effectuée.

- **Les mesures utilisées pour évaluer le système SRIDOP**

  L'évaluation des systèmes de recherche d'informations est généralement faite d'après des indicateurs classiques tels que le rappel et la précision. En recherche d'information, le taux de rappel mesure la quantité de documents pertinents relevés par rapport à la quantité totale de documents pertinents du corpus et le taux de précision mesure la quantité de documents pertinents parmi les documents retenus.

$$Rappel = \frac{Documents\ pertinentes\ sélectionnés}{Documents\ pertinentes}$$
$$Précision = \frac{Documents\ pertinentes\ sélectionnés}{Documents\ sélectionnés}$$

  Dans le cadre de la recherche d'informations, ces indicateurs permettent de mesurer la quantité d'informations retrouvées parmi l'information qui aurait dû être relevée. Une mesure supplémentaire (F-mesure) permet de faire une synthèse entre rappel et précision, en favorisant les systèmes dont les mesures de rappel et de précision sont homogènes. Elle est généralement établie comme suit :

$$F - Mesure = \frac{2 * Précision * Rappel}{Précision + Rappel}$$

  Afin de procéder à l'évaluation du système, nous définissons un certain nombre d'indicateurs. Pour chaque module du système, nous comparons le résultat fourni par le système au résultat fourni par un expert humain. Le résultat de la comparaison peut être une égalité (le système a fourni la même solution que l'expert), une inclusion (le système a fourni une réponse partielle) ou un échec (le système n'a pas fourni de réponse ou a fourni une réponse erronée).

- Résultats de l'évaluation du système SRIDOP

  SRIDOP est un système qui applique une analyse linguistique de surface pour l'annotation des objets pédagogiques et leur extraction selon le besoin de l'utilisateur. L'information obtenue par une analyse linguistique est évaluée en premier lieu. Le système est contraint de prendre une décision : il doit fournir, s'il y a lieu, au moins une indication pour le type de l'objet pédagogique repéré. Pour cela le système utilise la méthode d'exploration contextuelle.

Cette évaluation est faite sur plusieurs étapes suivant les modules développés.

- Evaluation de la conversion et de la segmentation
  L'évaluation de la phase de conversion donne un résultat très bon sans échec sur l'ensemble des documents traités et qui sont des fichiers de plusieurs types (HTML, PDF, DOC, PPT, etc.). Pour la phase de segmentation, les documents pédagogiques sont assez irréguliers et présentent généralement des erreurs de structuration. Ceci fait que le travail de segmentation ne peut atteindre 100 %. Nous avons testé notre segmenteur sur un corpus d'évaluation composé de 300 documents de différents formats (PDF, PPT, DOC, HTML, etc.) téléchargés à partir du Web et rassemblés auprès de plusieurs enseignants. Nous ne pouvons pas dire que le résultat de la segmentation est sans échec, mais tout de même nous avons obtenu un taux près de 90 %.
  Les erreurs de segmentation sont dues à certains problèmes rencontrés dans le corpus. Ce sont les cas des segments terminés par un sigle et qui commencent par un nom propre et le cas d'un segment qui se termine par une lettre majuscule. Aussi dans les corpus traitant de sujets mathématiques où l'utilisation des inconnus "X, Y, etc." est très courante et entraînent des taux d'échec plus considérables. Ces problèmes sont fréquents car les documents sont de nature pédagogique.

  Dans le cadre du module suivant (annoation des objets pédagogiques), nous optons pour une évaluation basée sur les mesures de Précision et de Rappel. Nous présentons les résultats de chacune de ces mesures séparément car les paramètres d'évaluation ne sont pas les mêmes.
- Evaluation de la précision du module d'annotation
  Pour la mesure de précision, nous avons confié la tâche d'évaluation de ce module à un groupe de trois personnes : étudiants et enseignants. Le nombre d'évaluateurs est fixé à un nombre impair (trois) pour pouvoir trancher sur la justesse du type affecté à l'objet pédagogique annoté. Chaque personne a la charge d'évaluer le corpus composé de 300 documents.
  La segmentation et l'annotation de notre corpus a abouti à l'obtention de **N** segments textuels, répartis en *Na* segments annotés et *Nn* segments non annotés. Pour un même corpus, chaque évaluateur doit effectuer la tâche en indiquant, pour chaque annotation, le nom du fichier, le contenu de l'objet annoté, le type pédagogique de l'objet annoté et son sous-type.

Il doit indiquer aussi pour chaque objet : Positif (le type pédagogique affecté à l'annotation est correcte), Négatif (le type pédagogique affecté à l'annotation est faux).

L'évaluateur peut proposer un type pédagogique à la place du type affecté incorrectement (cf. Tab.4.1). Un exemple de tableau rempli par un évaluateur est présenté dans l'annexe C.

Nom du fichier	**Objet annoté**	**Type pédagogique**	**Sous-Type**	**Positif**	**Négatif**
Bases_De_ Donnees_ relationnelles_et _contraintes _SQL.txt	Certaines contraintes sont le reflet du modèle de données et permettent d'assurer la cohérence fonctionnelle des relations entre les tables	Caractéristique	Signe Distinctif	*	

TABLEAU 4.1: Format de tableau remplit par l'évaluateur

D'après l'évaluation effectuée par les 3 évaluateurs, nous concluons qu'il ya 3 cas de partage du jugement :

- Des annotations où tous les évaluateurs n'étaient pas d'accord sur le type pédagogique affecté aux objets pédagogiques. En effet, dans plusieurs cas, certains évaluateurs jugent " Positif" une annotation, contrairement à d'autres qui la jugent "Négatif".
- Des annotations où tous les évaluateurs étaient d'accord sur le type pédagogique "Positif".
- Des annotations où tous les évaluateurs étaient d'accord sur le type pédagogique "Négatif".

Ainsi nous définissons un taux de jugement Tj afin de trancher sur l'acceptation du type de l'annotation.

$$T_j = \frac{Nombre\ desévaluateurs\ qui\ ont\ accepté\ l'évaluation}{Nombre\ total\ desévaluateurs}$$

Ce taux sera :

– Null (égale à zéro) : dans le cas où le type affecté à l'objet annoté est rejeté par tous les évaluateurs.

– Egale à 1 (=1) dans le cas où le type affecté à l'objet annoté est accepté

– Compris entre 0 et 1 dans le cas où il n'y a pas unanimité de jugement entre les évaluateurs. Dans ce cas, le taux Tj sera amené à 0 si Tj $<$ 0,5 et à 1 si Tj $>$ 0,5.

Nous présentons ci-dessous un tableau qui illustre ces trois cas :

L'interprétation des résultats des évaluateurs sera effectuée par type d'objet en se basant sur le paramètre Tj. L'évaluation sera effectuée selon

Segment annoté	Type d'annotation	Eval1	Eval2	Eval3	Tj
La commande ALTER TABLE permet aussi de modifier la structure d'une table (changer les spécifications de la clé primaire ou ajouter une contrainte unique à une colonne)	Caractéristique	1	1	1	1
Si le caractère lu est un opérateur (binaires ou unaires) : : ≪ Faire attention à la priorité ≫ a- Il faut dépiler tous les opérateurs de priorité supérieure ou égale et les rangerdans la chaîne ≪ExPost≫	Exercice	0	0	0	0
L'objectif étant donc de maîtriser le développement c'est-à-dire réduire les coûts et assurer la qualité grâce à l'utilisation : - de méthodes (façon de faire quelque chose, - d'outils (programmes, langages	Méthode	0	1	0	0.3
Il se spécifie à l'aide de la syntaxe suivante :SET CONSTRAINTS { ¡liste_de_contraintes¿ — ALL } [DEFERRED — IMMEDIATE] Il n'opère que pour les contraintes qui n'ont pas été définies en tant que NOT DEFFERABLE	Caractéristique	1	1	0	0.7

Tableau 4.2: Illustration des résultats des évaluateurs

les types des objets pédagogiques à savoir : Définition, Plan, Exercice, Exemple, Cours, Caractéristique et Méthode.

Type pédagogique	Nbre Obj Annotés	Nbre Obj Annotés Correctement	Taux de Précision
Caractéristique	1513	1421	93.91
Cours	80	78	97.5
Définition	1305	1262	96,7
Exemple	1054	925	87.7
Exercice	812	512	63.0
Méthode	350	240	68,57
Plan	20	10	50.0
Total	5134	4448	86,63%

Tableau 4.3: Résultats de la précision du module d'annotation

Le taux de précision est calculé selon cette équation :

$$Taux\ deprécision(d'un\ type) = \frac{Nombre\ d'objet\ annotés\ correctement}{Nombre\ d'objetannotés}$$

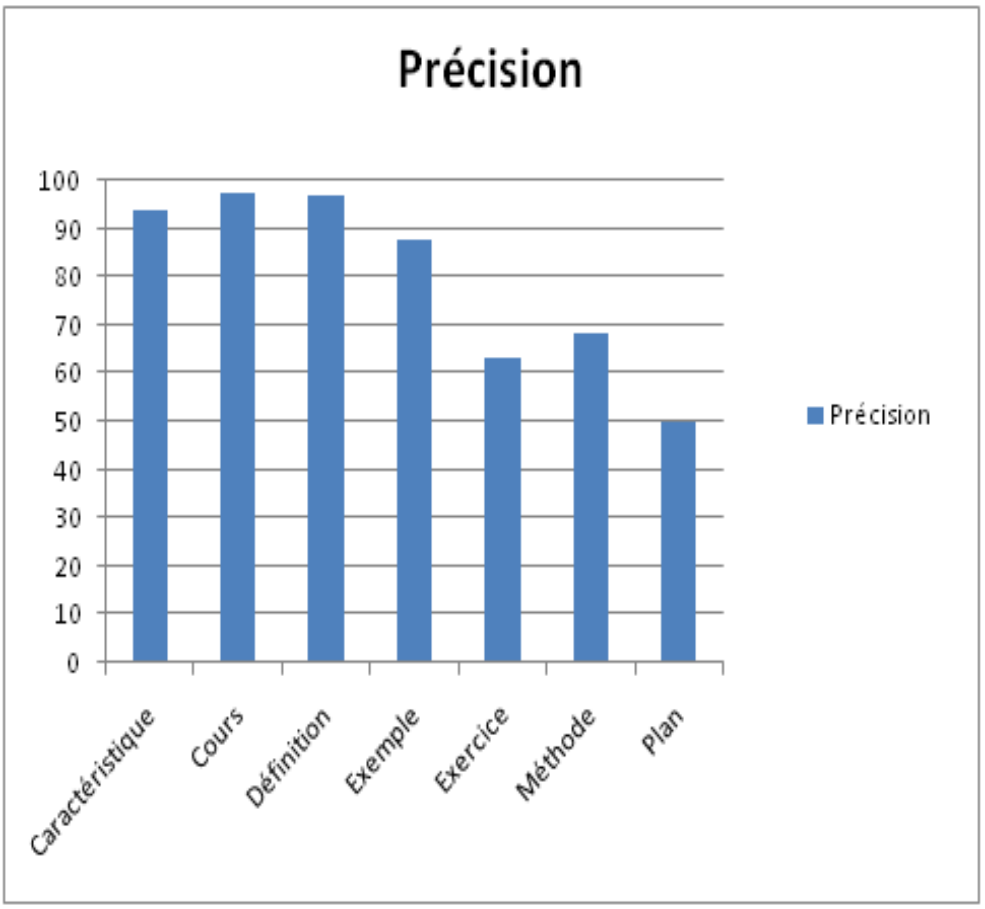


FIGURE 4.17: La mesure de précision des différents types pédagogiques dans le module d'annotation

La figure ci-dessus (cf. Fig.4.17) illustre les résultats obtenus (Précision) lors des tests de notre module d'annotation sur notre corpus d'évaluation. Ces résultats montrent que tous les types d'objets pédagogiques ne présentent pas les mêmes valeurs de précision. Nous remarquons que les deux types Exercice et Plan ont une précision plus faible que les autres. Nous présentons, dans ce qui suit, des exemples d'erreurs qui peuvent causer la faiblesse de la précision pour chacun des types pédagogiques.

- Définition

  L'objet pédagogique suivant était annoté incorrectement comme une Définition par notre système. La règle, qui a permis d'identifier l'objet pédagogique suivant est la règle RD1 dont l'indicateur est " sont définis". Dans ce cas, la règle n'est pas pertinente pour l'indicateur dans ce contexte : le type "Définition" n'est pas exprimé dans ce contexte.

Les utilisateurs anonymes sont définis par l'insertion de ligne avec User=" dans la table mysql.

- Exemple

  L'objet pédagogique suivant était annoté incorrectement comme un Exemple par notre système. La présence du mot "Exemple" a permis de déclencher la règle RE2 et d'annoter le segment textuel comme un Exemple, alors que c'était une simple énonciation du mot Exemple dans un plan de support de cours.

  > PLAN Chapitre 1 : Introduction au langage C 1 Introduction au langage C 2 Structure d'un programme C 3 Les expressions et les opérateurs en C 4 Les fonctions d'Entrées Sorties 5 Exemple Chapitre 2 : Les structures alternatives 1 Introduction 2 IF-ELSE 3 SWITCH Chapitre 3 : Les structures répétitives 1 Introduction 2 While 3 Do... While 4 For 5 Instructions de branchement inconditionnel Chapitre 4 : Les tableaux 1 Introduction 2 Les tableaux à une dimension 3 Les tableaux à deux dimensions

- Exercice

  Le verbe "Dire" est une occurrence de l'indicateur de la règle RX2. Sa présence déclenche donc l'annotation du segment textuel comme un "Exercice". Cependant, l'objet pédagogique suivant était annoté incorrectement comme un Exercice par notre système. En effet, la règle RX2 qui a permis d'identifier l'objet pédagogique, n'est pas pertinente pour l'indicateur "Dire" dans ce contexte : le type "Exercice" n'est pas exprimée dans ce contexte.

  > *Dire qu'un logiciel est de qualité sous entend qu'on puisse lui appliquer certains critères comme :*
  > *o L'adéquation aux besoins des utilisateurs,*
  > *o La fiabilité,*
  > *o L'efficacité,*
  > *o L'évolutivité.*

- Caractéristique

  L'objet pédagogique suivant était annoté incorrectement comme une Caractéristique par notre système. Dans ce cas aussi, la règle qui a permis d'identifier l'objet pédagogique, n'est pas pertinente pour l'indicateur dans ce contexte.

  > *Ce que je vous propose de visiter c'est comment SQL implémente les contraintes dans les bases de données relationnelles*

- Cours

  L'objet pédagogique suivant était annoté incorrectement comme un Cours par notre système. En effet, cet objet ne représente pas un support de cours mais c'est juste une énonciation de cours. En effet, la règle qui a

permis d'identifier l'objet pédagogique, est pas pertinente pour l'indicateur dans ce contexte : la notion n'est pas exprimée dans ce contexte.

> *Auto évaluation Cours n°1 LS1 UEO11 : Bases Mathématiques et Informatiques Nom : Prénom : Groupe de TD : Orientation : Temps : 12 min Barême : 1 point par réponse juste,* 0

- Plan

  Une simple détection du mot plan en dehors d'un titre d'un plan fait que l'objet annoté soit incorrectement annoté comme Plan.

- Méthode

  L'objet pédagogique suivant était annoté incorrectement comme une Méthode par notre système. En effet, c'est la définition du mot "Méthode" et non pas une expression linguistique du type "Méthode". Donc, la règle qui a permis d'identifier l'objet pédagogique, n'est pas pertinente pour l'indicateur dans ce contexte.

  > *Méthode : c'est un ensemble de concepts, de techniques d'aide à la résolution d'un problème*

Suite aux résultats de la première évaluation, nous avons apporté des améliorations aux règles d'EC et ce pour améliorer le taux de précision des résultats du module d'annotation.

– L'évaluation du rappel du module d'annotation

L'évaluation de notre système par la mesure du rappel des résultats obtenus est confiée à un évaluateur : un enseignant. Il a la charge d'évaluer un corpus composé de 30 documents. Sa tâche consiste à parcourir les fichiers annotés et de repérer les segments textuels non annotés et qui auraient dû être annotés, ainsi que leurs types pédagogiques. Suite à cette première évaluation, des corrections et des améliorations sont apportées aux règles d'annotation.

Type pédagogique	**Nbre Obj Annotés par le système**	**Nbre Obj Annotés par l'évaluateur**	**Taux de Rappel Rappel**
Caractéristique	135	176	76,7%
Cours	8	10	80%
Définition	170	212	80,18%
Exemple	123	162	75,46%
Exercice	81	90	90%
Méthode	32	40	80%
Plan	2	4	50%
Total	**551**	**694**	**65%**

Tableau 4.4: Résultats du rappel du module d'annotation

Le taux de rappel est calculé selon cette équation :

$$Taux\ de\ rappel(pour\ un\ type\ donné) = \frac{Nombre\ d'objet\ annotés\ par\ le\ système}{Nombre\ d'objet\ annotés\ par\ l'évaluateur}$$

La figure suivante (cf. Fig.4.18) illustre les résultats obtenus (Rappel) lors des tests de notre module d'annotation sur notre corpus d'évaluation. Ces résultats montrent que tous les types d'objets pédagogiques ne présentent pas les mêmes valeurs de Rappel.

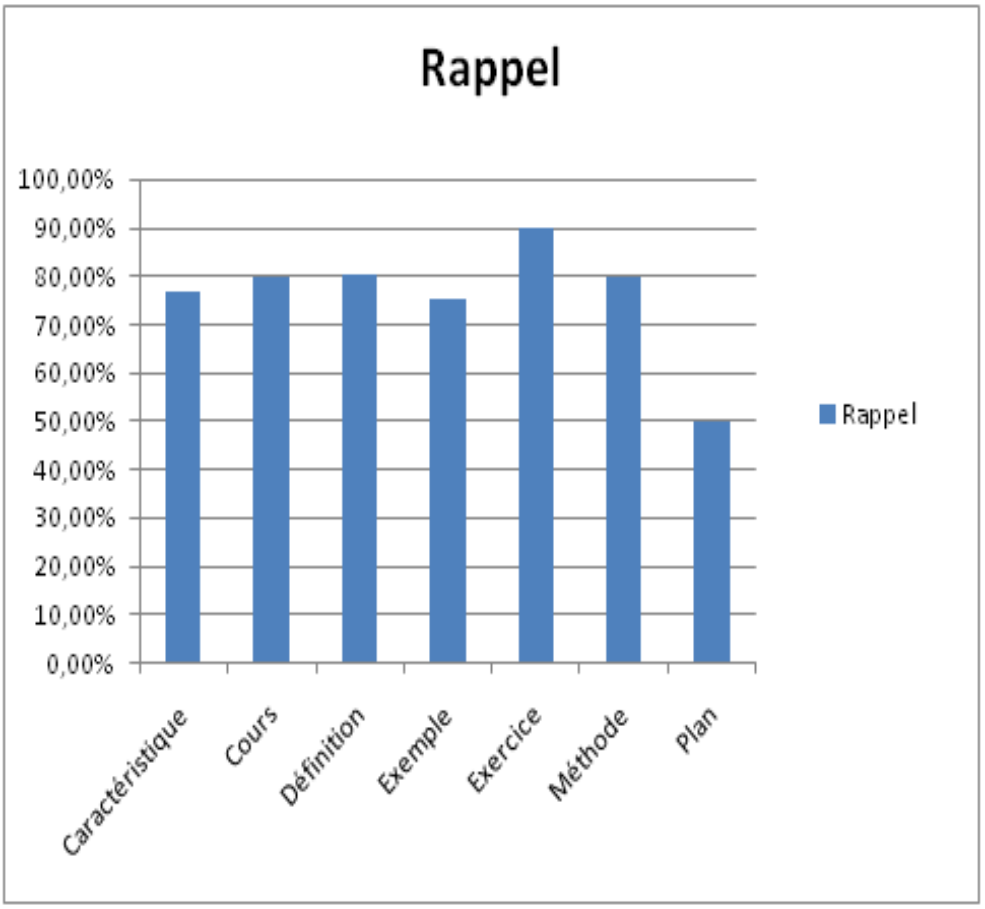


FIGURE 4.18: La mesure de rappel des différents types pédagogiques dans le module d'annotation

Nous présentons, dans ce qui suit, des exemples d'objets pédagogiques annotés par l'évaluateur. La non détection de ces derniers par notre système cause la faiblesse de la mesure de rappel pour chacun des types pédagogiques :

- Définition

  Les interprétations des connecteurs sont les mêmes que dans le calcul des propositions (elles portent sur les résultats d'interprétations de propositions, à valeurs dans {Faux, Vrai})

  Les marqueurs linguistiques exprimant cette définition "sont les mêmes que" ne faisaient pas partie de notre base de règles d'Exploration Contextuelle.

- Exercice

  A quelle vitesse percutera-t-elle le sol ? Les données (masse, hauteur, accélération terrestre) doivent être tirées de l'information (l'énoncé du problème) et de tes connaissances (la loi de Newton, l'accélération terrestre

  A quelle vitesse percutera-t-elle le sol ? Les données (masse, hauteur, accélération terrestre) doivent être tirées de Le marqueur linguistique "A quelle" ne fait partie de l'ensemble des marqueurs relatifs au type pédagogique "Exercice".

- Caractéristique

  Les trois caractéristiques fondamentales des objets sont : - L'état - Le comportement - L'identité L'état correspond aux valeurs instanciées de tous les attributs de l'objet

  Dans cet exemple, la présence du mot "trois" a empêché l'annotation du segment comme une caractéristique. En fait, ce mot ne fait pas partie de l'ensemble des règles.

- Cours

  Chapitre I : Introduction à l'Intelligence Artificielle

- Méthode

  La procédure est formulée en termes d'étapes très simples, du type : "si vous êtes dans l'état 42 et que le symbole contenu sur la case que vous regardez est '0', alors remplacer ce symbole par un '1', passer dans l'état 17, et regarder une case adjacente (droite ou gauche)"

  Suite à cette première évaluation, nous avons apporté des corrections à nos ressources linguistiques (Règles d'EC). Il s'ait d'exprimer les marqueurs linguistiques non repérés par notre système sous forme de règles d'Exploration Contextuelle. Ensuite, nous avons effectué une deuxième exécution et nous avons remarqué une nette amélioration dans les résultats obtenus.

∗ Evaluation de l'extraction des objets répondant à la requête utilisateur

Cette section présente l'évaluation de notre module de recherche des objets pédagogiques pertinents répondant à une requête utilisateur. Dans ce qui suit, nous commençons par la description du cadre d'évaluation avant de présenter les résultats de ces expérimentations.

- Cadre d'évaluation

  Le corpus que nous avons utilisé lors de cette phase est le même corpus utilisé lors de l'évaluation du module d'annotation des objets pédagogiques. C'est un corpus composé de 300 documents en français principalement de nature pédagogique. Ils ont une longueur moyenne de 10 pages.

  Afin d'évaluer la performance de notre système SRIDOP, nous proposons l'application de 35 requêtes avec les jugements de pertinence. Chaque requête est composée d'une première partie concernant le type de l'objet pédagogique recherché et une deuxième partie concernant le terme à rechercher. Pour chaque type de l'ensemble des 7 types pédagogiques, nous appliquons les mêmes 5 termes. Le tableau suivant présente un résumé des requêtes posées au système.

Type pédagogique	**Termes**			
Caractéristique	Génie Logiciel	Langage SQL	Intelligence Artificielle	Diagramme de classe
Cours	Génie Logiciel	Langage SQL	Intelligence Artificielle	Diagramme de classe
Définition	Génie Logiciel	Langage SQL	Intelligence Artificielle	Diagramme de classe
Exemple	Génie Logiciel	Langage SQL	Intelligence Artificielle	Diagramme de classe
Exercice	Génie Logiciel	Langage SQL	Intelligence Artificielle	Diagramme de classe
Méthode	Génie Logiciel	Langage SQL	Intelligence Artificielle	Diagramme de classe
Plan	Génie Logiciel	Langage SQL	Intelligence Artificielle	Diagramme de classe

Tableau 4.5: Résumé des réquêtes posées au système

Nous avons choisit de constituer des requêtes avec les mêmes termes pour chaque type d'objet pédagogique pour pouvoir préserver les mêmes paramètres d'évaluation et pour pouvoir comparer plus tard les résultats.

- Résultats expérimentaux

Un ensemble d'évaluations est effectué sur notre corpus d'évaluation. La pertinence des résultats obtenus est mesurée en termes de Rappel et Précision. Nous rappelons que, le taux de rappel mesure la quantité de documents pertinents relevés par rapport à la quantité totale de documents pertinents du corpus et le taux de précision mesure la quantité de documents pertinents parmi les documents retenus.

$$Rappel = \frac{Documents\ pertinentes\ sélectionnés}{Documents\ pertinentes}$$
$$Précision = \frac{Documents\ pertinentes\ sélectionnés}{Documents\ sélectionnés}$$

Dans le cadre de la recherche d'informations, ces indicateurs permettent de mesurer la quantité d'informations retrouvées parmi l'information qui aurait dû être relevée. Une mesure supplémentaire (F-mesure) permet de faire une synthèse entre rappel et précision, en favorisant les systèmes dont les mesures de rappel et de précision sont homogènes. Elle est généralement établie comme suit :

$$F - Mesure = \frac{2 * Précision * Rappel}{Précision + Rappel}$$

Nous avons confié la tâche d'évaluation de ce module à un seul évaluateur. Il a la charge d'évaluer le corpus composé de 300 documents. A travers une interface de recherche d'informations, il saisit les termes à rechercher, et choisit le type de l'objet pédagogique relatif au terme à rechercher. Les réponses aux requêtes sont affichées sous forme de liens permettant d'accéder à l'objet pédagogique répondant au besoin de l'utilisateur. Pour tester ce module de recherche d'objets pédagogiques, nous avons formulé les 35 requêtes présentés dans le tableau X1 présenté ci-dessus. Ces requêtes appartiennent aux différents domaines du corpus. Pour chaque type d'objet, nous avons illustré le nombre de réponses ramenées et le nombre de réponses jugées pertinentes compte tenu de l'ensemble des requêtes formulées. Les résultats sont résumés dans le tableau suivant (cf. Fig.4.6).

Type pédagogique	**Nbre Obj retournés par le système**	**Nbre Obj pertinents**	**Nbre Obj retournés par l'évaluateur**	**Précision**	**Rappel**	**F-Mesure**
Caractéristique	31	26	38	83,87%	68,42%	75,36%
Cours	6	5	5	83,33%	83,33%	83.33%
Définition	80	72	83	90%	86,74%	88,33%
Exemple	71	66	76	92,95%	86,84%	89,79%
Exercice	63	55	80	87,30%	68,75%	76,92%
Méthode	15	10	13	66,66%	76,92%	71,42%
Plan	4	3	5	75%	60%	66,66%

TABLEAU 4.6: Résultat de l'évaluation du module de recherche des objets pédagogiques

Nous illustrons ces valeurs relatives à chacun des types d'objets dans la figure ci-dessous (cf. Fig.4.19).

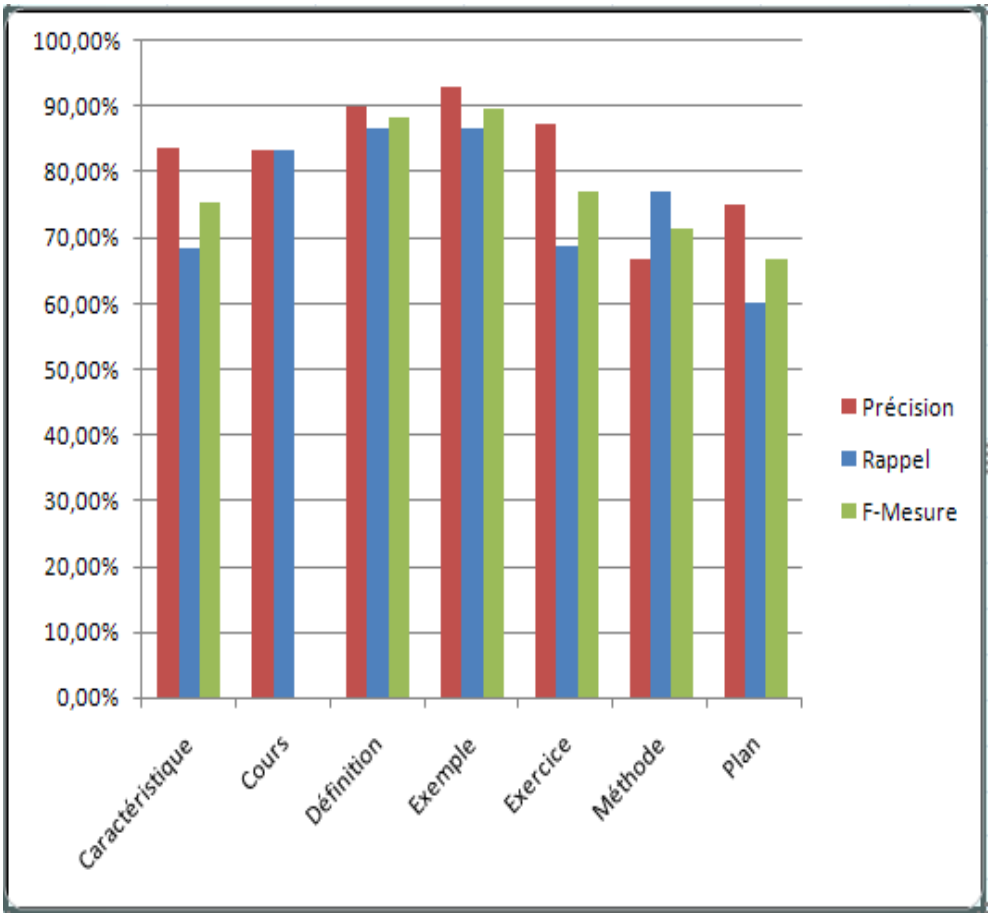


FIGURE 4.19: Précision, Rappel et F-Mesure de l'étape de classement des objets

La figure ci-dessus présente, pour chaque type d'objet (représenté sur l'axe des abscisses), sa valeur de précision représentée en bleu, sa valeur de rappel en pointillé et sa valeur de F-Mesure représentée en rayures. Nous constatons que les valeurs de précision sont comprises entre 75% et 87% et que celles du rappel entre 74% et 85%.

Durant la réalisation des évaluations des différents modules, nous avons remarqué que les résultats d'un module dépendent largement des résultats du module qui le précède. Ceci s'explique par le fait que les modules se succèdent dans leurs exécutions les uns après les autres tout en gardant leur indépendance.

· Evaluation comparative

Dans cette section, nous essayons de positionner la performance de notre méthode d'annotation et d'extraction des objets pédagogiques à partir de documents textuels par rapport aux résultats d'autres systèmes similaires au nôtre. Mais la tâche de comparaison s'annonce difficile En effet, il ya peu de systèmes qui traitent exactement les mêmes tâches que le nôtre. Et parmi ces systèmes, rares qui ont présenté leurs résultats d'évaluation. Nous avons choisi le système WebLearn (Liu et al., 2003) pour illustrer ses résultats concernant la recherche de la "Définition" de plusieurs termes en Anglais. D'après (Liu et al., 2003), les résultats sont les suivants (cf. Tab.4.7).

Les sujets de recherche	**Précision**
Artificial Intelligence	50.00
Data Mining	70.00
Web Mining	75.00
Machine Learning	77.78
Computer Vision	33.33
Retational Calculus	83.33
Linear Algebra	40.00
Neural Network	80.00
Fuzzy Logic	90.00
Time Series	50.00
Query Languages	20.00
Question Answering	75.00
Bioinformatics	60.00
DataBase Design	83.33
...	
Moyenne	**61.23**

TABLEAU 4.7: Résultats de l'évaluation du système WebLearn

Nous constatons que les résultats varient d'un sujet à un autre

4.5 Conclusion

Dans ce chapitre, nous avons présenté, dans une première partie, le principe général de notre système SRIDOP assurant l'annotation et l'extraction d'objets pédagogiques répondant à une requête utilisateur.
Ce système permet de : (a) acquérir une collection de documents pédagogiques, (b) appliquer le prétraitement nécessaire à cette collection, (c) annoter les objets pédagogiques en se basant sur les ressources linguistiques qui peuvent être (d) gérés à travers notre système, (e) indexer les objets pédagogiques annotés, (f) rechercher et extraire les objets pédagogiques répondant à une requête utilisateur, (g) constitution de fiches pédagogiques.
Dans une deuxième partie, nous avons évalué notre système SRIDOP. Pour les besoins de l'évaluation nous avons utilisé un corpus de documents principalement pédagogiques. Chaque module a été évalué séparément.
L'évaluation de la phase de segmentation donne de très bon résultat sur notre type de corpus.
Pour le module d'annotation des objets pédagogiques ainsi que le module d'indexation, nous avons effectué une évaluation pour chacun des types d'objets pédagogiques. Les résultats de ces évaluations ont montré que notre méthode d'annotation sémantique et indexation des objets pédagogiques donne de meilleurs résultats comparativement aux classiques.
Nous concluons ainsi que la recherche d'informations basée sur l'annotation sémantique permet d'améliorer la performance de la recherche d'informations pédagogiques.
Toutefois, comme nous pouvons s'y attendre, l'effet des résultats de segmentation

Bibliographie

Alrahabi M., Desclés J-P., Suh J., "Direct Reported Speech in Multilingual Texts : Automatic Annotation and Semantic Categorization", Actes de FLAIRS 2010, Florida, USA, 2010.

Alrahabi M., EXCOM-2 : plateforme d'annotation automatique de catégories sémantiques. Applications à la catégorisation des citations en français et en arabe, Thèse de doctorat, Université Paris-Sorbonne, 2010.

Atanassova I., Exploitation informatique des annotations sémantiques automatiques d'Excom pour la recherche d'informations et la navigation, Thèse de doctorat, Univesité Paris-Sorbonne, 2012.

Amardeilh F., Web sémantique et informatique linguistique : propositions méthodologiques et réalisation d'une plateforme logicielle, Thèse de doctorat en Informatique, Langage et Modélisation, Université Paris X-Nanterre, 2007.

Audibert L., Cours d'Informatique et linguistique II (INFZ24), Jeune équipe DELIC, Université de Provence, 2003.

Balpe J-P., Lelu A., Papy F., Saleh I., "Techniques avancées pour l'hypertexte", Ed. Hermès, ISBN 2-86601-522-3, 1996.

Baziz M., Indexation Contextuelle guidée par ontologie pour la recherche d'information, Thèse de Doctorat, Institut de Recherche en Informatique de Toulouse, 2005.

Bellot P., Méthodes de classification et de segmentation locales non supervisées pour la recherche documentaire, Thèse de Doctorat en Informatique, Université d'Avignon et des Pays de Vaucluse, 2000.

Ben Hazez S., Un modèle d'exploration contextuelle des textes : filtrage et structuration des informations textuelles, modélisation et réalisation informatique (Système Sémantex), Thèse de doctorat, Université de Paris-Sorbonne, 2002.

Berri J., Contribution à la méthode d'exploration contextuelle. Applications au résumé automatique et aux représentations temporelles. Réalisations du système SERAPHIN, Thèse de Doctorat, Université Paris-Sorbonne, 1996(a).

Berri J., "Mise en œuvre de la méthode d'exploration contextuelle pour le résumé automatique de textes. Implémentation du système SERAPHIN", Actes du colloque CLIM'96, Montréal, Canada, 1996(b).

Bertin G., Bertrand A., Bourda Y., et al. L'indexation des ressources pédagogiques numériques : un partenariat à créer entre les SCD et les services TICE au sein des universités. Lyon : école nationale supérieure des sciences de l'information et des bibliothèques, 2004, p. 87.

Bertin M., Desclés J.P., Djioua B., Krushkov Y.,"Automatic Annotation in Text for Bibliometrics Use", Actes de FLAIRS, 2006.

Bertin M., Bibliosémantique : une technique linguistique et informatique par exploration contextuelle, Thèse de doctorat, Université Paris-Sorbonne, 2011.

Bertrand-Gastaldy S., L'évolution de la gestion de l'information documentaire sous l'impulsion des nouvelles technologies, Dans Terminogramme, Bulletin d'information terminologique et linguistique, n°55, 1990, p. 25-31.

Blais A., Atanassova I., Desclés J.P., Zhang M., Zighem L., "Discourse Automatic Annotation of texts : An application to Summarization", Actes de FLAIRS, Florida, 2007.

Blaschke C., Andrade M.A., Ouzounis C., Valencia A., "Automatic Extraction of biological information from scientific text : protein-protein interactions", Actes de 7th International Conference on Intelligent Systems in Molecular Biology, (ISMB'99), Heidelberg, Germany, AAAI Press, 1999, p. 60-67.

Blumentritt R., Johnston R., "Toward a strategy for knowledge management", Technology Analysis and Strategic Management, 1999, p. 287-300.

Bodain Y., "Logiciel d'annotation pour la conception de cours sur le web sémantique", Actes de IHM, Montréal, 2006.

Boucetta Z., Boufaida Z., Yahiaoui L., "Appariement sémantique des documents à base d'ontologie pour le e-Recrutement", Actes de COSI, 2008, Tizi-Ouzou, Algérie.

Boughanem M., Les Systèmes de Recherche d'Information : d'un modèle classique à un modèle connexionniste, Thèse de Doctorat, Université Paul Sabatier, Toulouse (France), Décembre 1992.

Boughanem M., Contribution à la Formalisation et à la Spécification des Systèmes de Recherche et de Filtrage d'Information. Habilitation à Diriger les Recherches, Université Paul Sabatier de Toulouse, 2000.

Boughanem M., "Introduction à la Recherche d'Information", EARIA, 2006.

Boughanem M., Savoy J., Recherche d'information : état des lieux et perspectives, Hermès, Lavoisier, 2008, p. 342.

Bouhafs A., Utilisation de la méthode d'exploration contextuelle pour une extraction d'informations sur le Web dédiées à la veille, réalisation du système informatique JAVA Veille, Thèse de doctorat, Université Paris-Sorbonne, 2005.

Bousbia N., Balla A., "Processus de modélisation de contenus pédagogiques destinés à la FOAD ", Revue électronique des technologies de l'information, N° 3, 9 mai 2007.

Buckley C., Salton G., Allan J., "The effect of adding relevance information in a relevance feedback environment". Actes de International ACM SIGIR Conference, 1994, p.292-300.

Buffa M., Dehors S., Faron-Zucker C., Sander P., "Vers une approche Web Sémantique dans la conception d'un système d'apprentissage. Revue du projet TRIAL SOLUTION", Actes de WebLearn, 2005.

Catteau O., Le cycle de vie de l'Objet pédagogique et de ses métadonnées, Thèse de Doctorat, Université de Toulouse III-Paul Sabatier, 2008.

Cellier P., Ferré S., Ridoux O., and Ducasse M., "A parameterized algorithm to explore formal contexts with a taxonomy." International Journal of Foundations of Computer Science 19, no. 02, 2008 : 319-343.

Cernea D., Moral E., Gayo J.E., "SOAF : Semantic indexing system based on collaborative tagging", Interdisciplinary Journal of E-learning and Learning Objects, Vol. 4, 2008.

Chabert-Ranwez S., "Composition automatique de documents Hypermédia Adaptatifs à partir d'Ontologies et de Requêtes intentionnelles de l'Utilisateur", Thèse de Doctorat, Université de Montpellier II, 2000.

Chagnoux M., Ben Hazez S., Desclés J-P.,"Identification automatique des valeurs temporelles dans les textes ", Actes de 10ème conférence sur le traitement automatique des langues (TALN' 2003), Batzsur-mer, 2003.

Chai H.,"Automatic Annotation for Korean-Approach based on the Contextual Exploration Method", Database and Expert Systems Applications, 2007, p. 278 - 282.

Christiansen J.-A., Anderson T., “Feasibility of course development based on learning objects : Research analysis of three case studies ”, International Journal of Instructional Technology and Distance Learning, Vol. 1, n° 3, 2004.

Christos F., Information Retrieval : Data Structures and Algorithms, chapter Signature Files, Prentice Hall, Upper Saddle River, New Jersey, 1992, p. 44-65.

Claveau V., Tavenard R., Amsaleg L., “Vectorisation des processus d’appariement document-requête”, Actes de CORIA, 2010.

Cleuziou G.,“Regroupements non-disjoints de mots pour la classification de documents”, Actes de CORIA, 2004.

Cleuziou G., Dias, G., “Apprentissage de mesures de similarité sémantiques : étude d’une variante de la mesure InfoSimba”, The first Joint Meeting of the Société francophone de classification and the Classification and Data Analysis Group of the Italian Society of Statistics (SFC-CLADAG 2008), Italy, 2008, p. 233-236.

Crispino G., Ben Hazez S., Minel J-L., “Architecture logicielle ContextO, plateforme d’ingénierie linguistique filtext”, Conférence sur le traitement automatique des langues (TALN), 1999, p. 327-332.

Daoust F., Marcoux Y.,“Logiciels d’analyse textuelle : vers un format XML-TEI pour l’échange de corpus annoté”, Actes de JADT, 2006.

Dehors S., Faron-Zucker C., Stromboni J.P., Giboin A.,“Des annotations Sémantiques pour apprendre : l’Expérimentation QBLS”, Actes de WebLearn, 2005.

Delestre N., METADYNE : Un Hypermédia Adaptatif Dynamique pour l’enseignement, Thèse de Doctorat, Université de Rouen, 2000.

Denoue L., Vignollet L., “Yawas : un outil d’annotation pour les navigateurs du web”, Actes de IHM, 1999.

Desclés J.P., Jouis C., Oh H-G., Reppert D.,“Exploration Contextuelle et sémantique : un système expert qui trouve les valeurs sémantiques des temps de l’indicatif dans un texte”, Actes de Knowledge modeling and expertise transfert, D. Hérin-Aime, R. Dieng, J-P. Regourd, J-P. Angoujard (eds.), IOS Press, 1991.

Desclés J.P., Jouis C., “L’exploration contextuelle : une méthode linguistique et informatique pour l’analyse automatique de textes”, ILN, Nantes, 1993.

Desclés J.P.,“Système d’exploration Contextuelle”, Co-texte et calcul du sens, Caen, 1997, p. 215-232.

Desclés J. P., “Contextual Exploration Processing for Discourse Automatic Annotations of Texts”, FLAIRS, 2006.

Desclés J.P., Djioua B.,“La recherche d’informations par accès aux contenus sémantiques : vers une nouvelle classe de systèmes de recherche d’informations et de moteurs de recherche (Aspects linguistiques et stratégiques)”. Revue Roumaine de Linguistique, Tome LII, N° 1-2, 2007.

Desclés J-P., Le Priol F., Annotations automatiques et recherche d’informations, Hermès - Traite IC2 – série Cognition et Traitement de l’information, 2009.

Desclés J., Makkaoui O., Desclés J-P,“Towards automatic thematic sheets based on discursive categories in biomedical litterature”, Actes de International Conference on Web Mining and Semantics (WIMS’11), Sogndal, Norway, 2011.

Desmontils E., Jacquin C.,“Annotation sur le web : notes de lecture”, Journées de l’AS Web sémantique, France, 2002.

Desmontils E., Jacquin C., Simon L., Vers un système d’annotation distribué, Rapport de recherche à l’Institut de Recherche en Informatique de Nantes, Février 2003.

Desmoulins C., Grandbastien M., “Des ontologies pour indexer des documents techniques pour la formation professionnelle”, Actes de la conférence Journées Francophones d’Ingénierie des connaissances IC’2000, Toulouse, 2000.

Dgim H., Smei H., Ben Hamadou A.,“Interprétation sémantique de requêtes utilisateurs par l’usage d’ontologies”, GEI, Hammamet, Tunisie, 2006.

Dingli A., Ciravegna F., Wilks Y., ”Automatic Semantic Annotation using Unsupervised Information Extraction and Integration”, International Conference on Knowledge Capture (K-CAP), 2003.

Dinh D., Tamime L.,“Vers un modèle d’indexation sémantique adapté aux dossiers médicaux de patients”. Actes de CORIA, 2010.

Djioua B., Garcia-Flores J., Blais A., Desclés J.P., Guibert G., Jackiewe A., Le Priol F., Nait-Baha L., Sauzay B., “EXCOM : an automatic annotation engine for semantic information”, Actes de FLAIRS, Florida, 2006.

Djioua B., Desclés J.P., “Indexing documents by Discourse and semantic contents from automatic Annotations of Texts”, Actes de FLAIRS, Florida, 2007.

Duval E., “An Open Infrastructure for Learning - the ARIADNE project - Share and Reuse without boundaries”, Actes de ENABLE 99 : Enabling Network-Based Learning, Espoo, Finland, 1999.

Elkhlifi A., Faiz R., “French-Written Event Extraction Based on Contextual Exploration”, Actes de FLAIRS, AAAI Press, Florida, 2010.

Faiz R., “Identifying relevant sentences in new articles for event information extraction”, International Journal of Computer Processing of Oriental Languages, Vol.19, No.1, 2006.

Faiz R., Smine B., Desclés J.P., “Relevant Learning Objects Extraction Based on Semantic Annotation of Documents”, International Conference on Web Intelligence, Mining and Semantics (WIMS’12), ACM, Juin 2012, Craiova, Romania.

Flory L., “Les caractéristiques d’une ressource pédagogique et les besoins d’indexation qui en résultent”, Journée d’étude sur l’Indexation des ressources pédagogiques numériques, Ennsib, Villeurbanne, 2004.

Fort K., Les ressources annotées, un enjeu pour l’analyse de contenu : vers une méthodologie de l’annotation manuelle de corpus. Thèse de Doctorat en Informatique, Université Paris 13, 2012.

Garcia D., Analyse automatique des textes pour l’organisation causale des actions. Réalisation du système automatique Coatis, Thèse de doctorat, Université Paris-Sorbonne, 1998.

Garcia-Flores J., Annotation sémantique des spécifications informatiques de besoins par la méthode d’Exploration Contextuelle : une contribution des méthodes linguistiques aux conceptions de logiciels, Thèse de doctorat, Université Paris-Sorbonne, 2007.

Gibbon D., Moore R., Winski R., Handbook of Standards and Resources for Spoken Language Systems, Vol 4, Spoken Language Reference Materials, XVI, 1998, p. 242.

Gillard L., Indexation de documents annotés, Rapport de Mémoire de Mastère, Faculté des sciences de Luminy, Université de la méditerranée, 2002.

Goker A., Davies J., Information Retrieval, Searching in the 21st Century, Ed. Wiley, Editeurs Goker et Davies, 2009, p. 295.

Gonnet G., Baeza-Yates R., Snider T., Information Retrieval : Data Structures and Algorithms, chapter New Indices for Text : PAT Trees and PAT Arrays, Prentice Hall, Upper Saddle River, New Jersey, 1992, p. 66-82. Greenwood M.A., Saggion H., “A Pattern Based Approach to Answering Factoid, List and Definition Questions”, Actes de RIAO 2004, Avignon, France, 2004.

Habert B., Nazarenko A., Salem A., Les linguistiques de corpus. Collection U, série “Linguistique”, Armand Colin, Paris, 1997, p. 240.

Haddad H., Berrut C., Bruandet M.F., “Un modèle vectoriel de recherché d’informations adapté aux documents vidéo”, Actes de CORESA, 1996.

Hamdi S., Lopes Gancarski A., Bouzeghoub A., and Ben Yahia S., “Enriching ontologies from folksonomies for Elearning : DBpedia case.” In Advanced Learning Technologies (ICALT), 2012 IEEE 12th International Conference on, pp. 293-297. IEEE, 2012.

Hamdi S., Lopes Gancarski A., Bouzeghoub A., and Ben Yahia S., “Enriching the DBpedia ontology with shared conceptualizations from folksonomies.” In Computer Software and Applications Conference (COMPSAC), 2012 IEEE 36th Annual, pp. 551-556. IEEE, 2012.

Handschuh S., Staab S., Ciravegna F.,“S-CREAM- Semi-automatic CREAtion of Metadata. Lecture Notes”, Computer Science, Vol. 2473, 2002, p. 358-372.

Handschuh S., Staab S., Maedche A., “CREAM- Creating relational metadata with a component-based, ontology-driven annotation framework”, First International Conference on Knowledge Capture (K-CAP), 2001.

Harman D., Baeza-Yates R., Fox E., Lee W., Information Retrieval : Data Structures and Algorithms, chapter Inverted Files, Prentice Hall, Upper Saddle River, New Jersey, 1992, p. 28-43.

Hassan S., Mihalcea R., “Learning to identify educational materials”. Actes de RANLP, Bulgaria, 2009.

Henze N., Nejdl W., Wolpers M.,“ Modelling constructivist teaching functionality and structure in the KBS hyperbook system”, Actes de AI-ED 99 Workshop on Ontologies for Intelligent Educational Systems, Le Mans, France, 1999.

Jackiewicz A., L’expression de la causalité dans les textes, contribution au filtrage sémantique par une méthode informatique d’exploration contextuelle. Thèse de doctorat, Université Paris-Sorbonne, 1998.

Jacquemin C., Zweigenbaum P., Traitement Automatique des Langues pour l’accès au contenu des documents, Le document Multimédia en Sciences du Traitement de l’Information, CÉPADUÈS-Éditions, Toulouse, 2000, p. 71-109.

Jelassi M. N., Ben Yahia S., and Mephu Nguifo E., “A personalized recommender system based on users’ information in folksonomies.” In Proceedings of the 22nd International Conference on World Wide Web, pp. 1215-1224. ACM, 2013.

Jouis C., Contribution à la conceptualisation et à la modélisation des connaissances à partir d'une analyse linguistique de textes, Thèse de doctorat, EHESS, Paris, 1993.

Kachroudi M., Ben Moussa E., Zghal S., and Ben Yahia S., "Ldoa results for oaei 2011." In Proceedings of the 6th International Conference on Ontology Matching-Volume 814, pp. 148-155. CEUR-WS. org, 2011.

Kachroudi M., Hassen W., Zghal S., and Ben Yahia S., "Large Ontologies Partitioning for Alignment Techniques Scaling." In WEBIST, pp. 165-168. 2013.

Kachroudi M., Zghal S., and Ben Yahia S., "Using linguistic resource for cross-lingual ontology alignment." International Journal of Recent Contributions from Engineering, Science & IT (iJES) 1, no. 1, 2013 : 21-27.

Kachroudi M., Zghal S., and Ben Yahia S., "OntoPart : at the cross-roads of ontology partitioning and scalable ontology alignment systems." International Journal of Metadata, Semantics and Ontologies 8, no. 3, 2013 : 215-225.

Kachroudi M., Zghal S., and Ben Yahia S., "Paramétrage intelligent de l'alignement d'ontologies par l'intégrale de Choquet." In EGC, pp. 377-382. 2013.

Kahan J., Koivunen M.R., Prud'Hommeaux E., Ralph R.,"Annotea : An Open RDF Infrastrucutre for Shared Web Annotations", The 10th International Conference on World Wide Web, Hong Kong, 2001.

Kaiser T. B., Schmidt S. E., and Joslyn C. A., "Adjusting annotated taxonomies." International Journal of Foundations of Computer Science 19, no. 02, 2008 : 345-358.

Kamoun K., and Ben Yahia S., "A novel global measure approach based on ontology spectrum to evaluate ontology enrichment." complexity 39, no. 17, 2012.

Katz B.,"Annotating the World Wide Web Using Natural Language", Actes de the 5th RIAO Conference on Computer Assisted Information Searching on the Internet (RIAO '97), 1997.

Kiryakov A., Popov B., Terziev I., Manov D., Ognyanoff D., "Semantic Annotation, Indexing and Retrieval", Journal on web semantics, 2004.

Laublet Ph., Reynaud Ch., Charlet J., "Sur quelques aspects du web sémantique", Actes des deuxièmes assises du GDR 13, 2002, p. 59-78.

Le Priol F., Extraction et capitalisation automatiques de connaissances à partir de documents textuels. SEEK-JAVA : identification et interprétation de relations entre concepts, Thèse de doctorat, Université Paris-Sorbonne, 2000.

Le Priol F., Blais A., Desclés J.P., Djioua B., Garcia-Flores J., Guibert G., Jackiewicz A., Nait-Baha L., Sauzay B., "Automatic annotation of localization and identification relations in plateform EXCOM", Actes de FLAIRS, 2006.

Le Priol F., "Automatic Annotation of Discourse and Semantic Relations supplemented by Terminology Extraction for Domain Ontology Building and Information Retrieval", Actes de FLAIRS, 2007.

Lee M., Tsai K., Wang T.,"A practical ontology query expansion algorithm for semantic-aware learning objects retrieval", Computer & Education, Vol. 50, 2008 p. 1240-1257.

Leech G., Corpus annotation : Linguistic information from computer text corpora, Chapitre Introducing corpus annotation, Longman, London, p. 1-18, 1997.

Liu B., Chin C.W., Ng H.T.,"Mining topic-specific concepts and definitions on the web", Actes de 12th International World Wide Web Conference, ACM Press, New York, 2003, p.251-260.

Lowe J.B, Baker C.F, Fillmore C.J.,"A Frame-Semantic Approach to Semantic Annotation", Actes de the SIGLEX Workshop on Tagging Text With Lexical Semantics : Why, What, and How ? Washington, USA, 1997.

Luhn H., "The automatic creation of literature abstracts", IBM Journal of Research and Development, Vol. 2, n° 2, 1958, p.159-165.

Makkaoui O., Desclés J., Desclés J-P., "Evaluation and performance improvement of the BioExcom system for the automativ detectyion of speculation in biomedical texts", Actes de Third Workshop on Building and Evaluating Resources for Biomedical Text Mining (BioTxtM 2012), LREC2012, Istanbul, Turkey, 2012.

Maron M., Kuhns J., "On relevance, probabilistic indexing and information retrieval", Journal of the Association for Computing Machinery, vol. 7, 1960, p. 216-244.

Marquié G.,"Eléments de compte rendu de la synthèse de Françoise Clerc", Colloque Aider et accompagner les élèves dans et hors l'école, CRAP/Cahiers pédagogiques, 2010.

Memmi D., Le modèle vectoriel pour le traitement de documents, Les cahiers du laboratoire Leibniz, IMAG-Genoble, France, n°14, 2000.

Meyer M., Rensing Ch., Steinmetz R.,"Categorizing Learning objects based on Wikipedia as Substitute Corpus", Actes de The first International Workshop on Learning Object Discovery & Exchange, 2007.

Michel C., Rouissi S., “Etude de l’organisation et caractérisation de l’information pédagogique pour construire des documents hypermédias adaptatifs diffusés sur le Web”, 5ème Colloque International sur le Document Électronique (CIDE’05), Inria (Ed.), Hammamet, Tunisie, 2002, p. 153-167.

Michelson M., Knoblock G.,“An Automatic Approach to Semantic Annotation of Unstructured, Ungrammatical Sources : A First Look”, AND, 2007.

Mille D., Modèles et outils logiciels pour l’annotation sémantique de documents pédagogiques, Thèse de Doctorat en Informatique, Université Joseph Fourier, Grenoble, 2005.

Minel J-L., Filtrage sémantique, Hermès, Paris, 2002.

Mitchell T.,“Machine learning”, New York : McGraw Hill, 1997.

Morizio C., La recherche d’information. Ed. Armand Colin, 2006, p.126.

Mourad G., Analyse informatique de signes typographiques pour la segmentation de textes et l’extraction automatique des citations. Réalisation des applications informatiques : Segatex et CitaRE, Thèse de Doctorat, Université Paris-Sorbonne, 2001.

Mourad G., “La segmentation de textes par exploration contextuelle automatique, présentation du module SegATex”, Inscription Spatiale du Langage : structure et processus ISLsp., Toulouse, 2002, IRIT, Université Paul Sabatier.

Muresan S., Klavans J.L.,“A Method for Automatically Building and Evaluating Dictionary Resources”, Actes de LREC, 2002.

Murray T., Madison A.W., Westall J.M.,“Colored Page Stealing”, Actes de 34th Annual ACM SouthEast Conference, 1996, p. 28-34.

Nanard M., Nanard J., “MacWeb : un outil pour élaborer des documents”, WOODMAN’89, Workshop on Object Oriented Document Manipulation, Rennes, 29-31 mai 1989.

Nestorov S., Ullman J.D., Wiener J.L., Chawathe S.S., “Representative objects : Concise representations of semistructured, hierarchial data”. Actes de International Conference on Data Engineering (ICDE’97), Birmingham, United Kingdom, IEEE Computer Society, 1997, p. 79-90.

Palmer D., Hearst A., “Adaptative sentence boundary disambiguation”, Actes de Conference on Applied Natural Language Processing, Stuttgart, Germany, 1994.

Paquet Ph., De l'information à la connaissance, Cahier de recherche, Numéro 2006-01 Laboratoire Orléanais de gestion (EA2635), Faculté de Droit, d'économie et de Gestion, 2006.

Paquette G., "L'ingénierie pédagogique à base d'objets et le référencement par les compétences", International Journal of Technologies in Higher Education, Vol.1, n° 3, 2004.

Pernin J.P.,"Objets pédagogiques : unités d'apprentissage activités ou ressources ?" Revue Sciences et Techniques Educatives, Hors série Ressources numériques, XML et éducation, Ed. Hermès, 2003, p.179-210.

Pillet V., Méthodologie d'extraction automatique d'information à partir de la littérature scientifique en vue d'alimenter un nouveau système d'information, Thèse de doctorat, Université de Droit, d'économie et des sciences d'Aix-Marseille, 2000.

Poibeau T., Extraction d'informations à base de connaissances hybrides, Thèse de doctorat, Université de Paris-Nord, 2002.

Prié Y., Garlatti S., "Méta-données et annotations dans le web sémantique", Dans Le web sémantique. Hors série de la revue Information-Interaction-Intelligence, Cépaduès, Toulouse, 2004, p. 45-68.

Robertson S. "The probability ranking principle in IR", Journal of Documentation, vol. 33, n°4, 1977, p. 294-304.

Rocchio J., "Relevance feedback information retrieval", In Gerard Salton, Ed.The Smart retrieval system experiments in automatic document processing, Prentice-Hall, Englewood Cliffs, NJ, 1971, p. 313-323.

Roussey S., Calabretto S., Pinon J-M., "SyDOM : un outil d'annotation pour le web sémantique", Journées Scientifiques du Web sémantique, 2002.

Salton G., The SMART retrieval system : Experiments in automatic document processing, Prentice Hall, 1970.

Salton G., "A Comparison between manual and automatic indexing methods", Journal of the American Documentation, Vol. 20, n°1, 1971, p. 61-71.

Salton G., Wong A., Yang C. S., "A vector Space Model for Automatic Indexing", Information Retrieval and language processing, 1975, p. 613-620.

Salton G., Mac Gill M.J., Introduction to modern Information Retrieval, New York MC Grew Hill Edition, 1983.

Salton G., “Automatic Text Indexing Using Complex Identifiers”, ACM Conference on Document Processing Systems, New Mexico, 1988.

Salton G., Buckley C., “Automatic structuring and retrieval of large text files”. Actes de the ACM 37 (2), 1994, p. 97-107.

Silberztein M., Dictionnaires électroniques et analyse automatique de textes, Le système INTEX, Paris, Masson, 1993.

Sinclair J., Preliminary Recommendations on Corpus Typology. EAGLES Document EAG-TCWG-CTYP/P, 1996.

Smei H., Ben Hamadou A., “Un système à base de métadonnées pour la création d’un cache communautaire-Cas de la communauté pédagogique”, IEBC, Hammamet, Tunisie, Juin 2005.

Smine B., Faiz R., Desclés J.P., “Analyse de documents pédagogiques en vue de leur annotation”, Revue des Nouvelles Technologies de l’Information (RNTI), Ed. Cépaduès, Janvier 2010, 429-434, 2010(a).

Smine B., Faiz R., Desclés J.P., “Pedagogical objects annotation based on Contextual Exploration”. Actes de The International Arabe Conference on Information Technologie, Benghazi, Lybie, 2010(b).

Smine, B., Faiz, R., “Automatic processing of learning objects for user’s query answering”. Actes de The International Conference on Information Technologie and e-services, Sousse, Tunisie, p. 297-302, 2011(a).

Smine B., Faiz R., Desclés J.P., “Extraction d’Informations pédagogiques pertinentes à partir de Documents Textuels”. Actes de Traitement Automatique des Langues Naturelles (TALN 2011), Montpellier, France, 2011(b).

Smine B., Faiz R., Desclés J-P., “The SRIDoP System Using Semantic Metadata for Web Database Processing”, Journal of Information Technology Review, Vol. 2, No. 3, 2011(c), p. 133-141.

Smine B., Faiz R., Desclés J-P., “Annotation et extraction automatique d’information pédagogiques à partir de documents textuels”, Actes du colloque International sur le document électronique (CIDE14), Rabat, Maghreb, 2011(d).

Smine B., Faiz R., Desclés J.P., “A semantic annotation model for indexing and retrieving learning objects”, Journal of Digital Information Management (JDIM), Vol. 9, No. 4, 2011(e), p. 159-166.

Smine B., Faiz R., Desclés J.P., “Extracting Relevant Learning Objects using a Semantic Annotation Method”, International Conference on Education and E-learning Innovations (ICEELI’2012), IEEE, 2012, Sousse, Tunisia.

Smine, B., Faiz, R., Desclés, J-P. “Relevant learning objects extraction based on semantic annotation”, International Journal of Metadata, Semantics and Ontologies, Vol. 8, No. 1, 2013, p.13-27.

Stapley B.J., Benoit G., “Bibliometrics : Information Retrieval and Visualization from Co-occurrence of Gene Names in MedLine Abstracts”, Actes de Pacific Symposium on Biocomputing (PSB’2000), vol. 5, Honolulu, 2000, p. 529-540.

Teissedre Ch., La définition, Etude linguistique, Utilisation dans un système de Recherche d’Information, Comparaison avec le “Define ” de Google. Master1-Informatique et Ingénierie de la langue pour la gestion de l’Information. Université de la Sorbonne, Paris IV, France, 2007.

Thomas J., Milward D., Ouzounis C., Pulman S., Caroll M., “Automatic Extraction of Protein Interactions from Scientific Abstracts”, Actes de Pacific Symposium on Biocomputing (PSB’2000), vol.5, Honolulu, 2000, p. 502-513.

Thompson C., Smarr J., Nguyen H., Manning C.,“Finding educational resources on the web : Exploiting automatic extraction of metadata”, Actes de ECML, Workshop on Adaptive Text Extraction and Mining, 2003.

Trabelsi C., Ben Jrad A., and Ben Yahia S., “Bridging folksonomies and domain ontologies : Getting out non-taxonomic relations.” In Data Mining Workshops (ICDMW), 2010 IEEE International Conference on, pp. 369-379. IEEE, 2010.

Trabelsi C., Nader J., and Ben Yahia S., “Scalable mining of frequent tri-concepts from folksonomies.” In Pacific-Asia Conference on Knowledge Discovery and Data Mining, pp. 231-242. Springer, Berlin, Heidelberg, 2012.

Vargas-Vera M., Motta E., Domingue J., Lanzoni M., Stutt A., Ciravegna F., “MnM : Ontology Driven Semi-Automatic and Automatic Support for Semantic Markup”, Journal Knowledge Acquisition, Modeling and Management, 2002.

Verbert K., Jovanovic J., Gasevic D., Duval E., “Repurposing Learning Object Components”, Actes de the OTM 2005 Workshop on Ontologies, Semantics and E-Learning (WOSE’05), Agia Napa, Cyprus, 2005.

Verney R., Annotation sémantique des services web. Master2- Image, Informatique, Ingénierie, Systèmes d’information. Université de Bourgogne, France, 2008.

Westerhout E., Monachesi P., “Creating glossaries using pattern-based and machine learning techniques”, Actes de Map of Language Resources, Technologies and Evaluation, 2008.

Wiley D. A., Learning object design and sequencing theory. Dissertation doctorale non publiée, Université de Brigham Young, Juin 2000.

Zghal S., Ben Yahia S., Mephu Nguifo E., and Slimani Y., “SODA : an OWL-DL based ontology matching system.” In Proceedings of the 2nd International Conference on Ontology Matching-Volume 304, pp. 261-267. CEUR-WS. org, 2007.

Zghal S., Ben Yahia S., Mephu Nguifo E., and Slimani Y., “SODA : Une approche structurelle pour l ?alignement d ?ontologies OWL-DL.” ZFO, 2007.

Zghal S., Kachroudi M., Ben Yahia S., and Mephu Nguifo E., “OACAS : results for OAEI 2011.” In Proceedings of the 6th International Conference on Ontology Matching-Volume 814, pp. 190-196. CEUR-WS. org, 2011.

Zghal S., Kamoun K., Ben Yahia S., Mephu Nguifo E., and Slimani Y., “EDOLA : Une nouvelle méthode d’alignement d’ontologies OWL-Lite.” In CORIA, pp. 351-367. 2007.

Zghal S., Mephu Nguifo E., Kamoun K., Ben Yahia S., and Slimani S., “A new alignment method for OWL-Lite ontologies using propagation of similarity over the graph.” In Database and Expert Systems Applications, 2007. DEXA’07. 18th International Workshop on, pp. 524-528. IEEE, 2007.

Zghal S., “Contributions à l’alignement d’ontologies OWL par agrégation de similarités.” PhD diss., Artois, 2010.

Annexe A

Documents pédagogiques

Document pédagogique N°1 :

ACCIDENT DE TRAVAIL

I. DEFINITIONS :

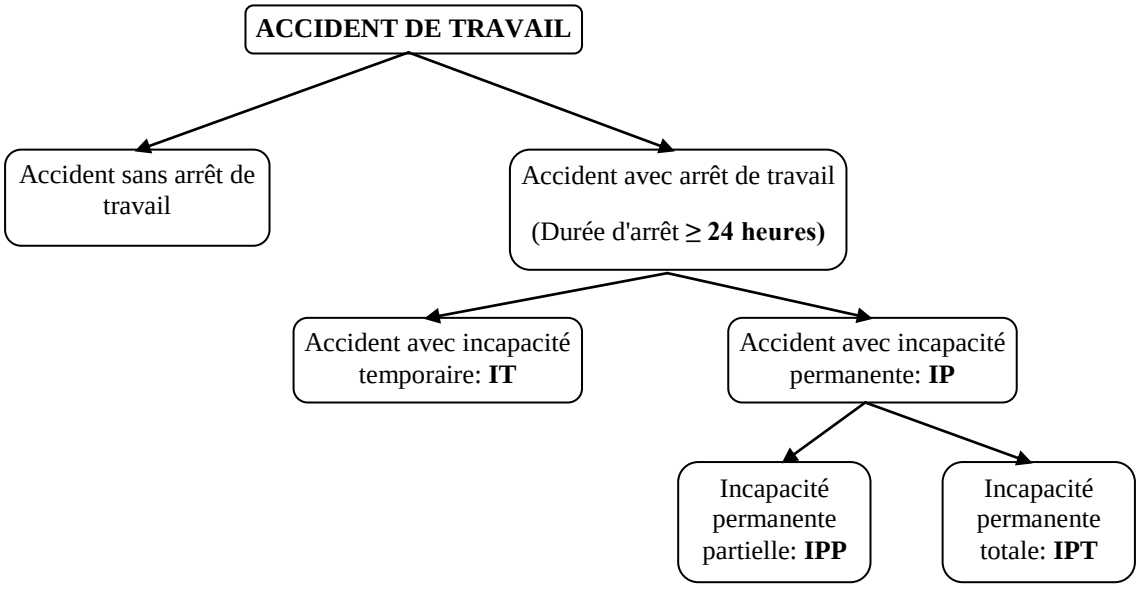
1. D'après le code de la sécurité sociale : Un accident de travail est considéré comme accident de travail, quelle qu'en soit la cause, l'accident survenu par le ***fait*** ou ***à l'occasion du travail*** à toute personne salariée ou travaillant dans le ***lieu de travail*** et pendant le ***trajet d'aller ou de retour*** entre :

- Sa résidence et le lieu de travail
- Le lieu de travail et le restaurant, la cantine où d'une manière plus générale, le lieu où le travailleur prend habituellement ses repas.

2. D'après la cour de cassation : L'accident de travail est légalement caractérisé par ***l'occasion violante*** et ***soudaine*** d'une cause extérieure provocant, au cours du travail, ***une lésion corporelle***.

3. Outre les accidents du travail définis comme nous venons de le voir, les travailleurs sont exposés à contracter certaines ***maladies*** et ***perturbations*** résultant directement de l'exercice de leur profession.

II. DIFFERENTS TYPES D'ACCIDENTS DU TRAVAIL:



III. LES MALADIES PROFESSIONNELLES:

Une maladie est «professionnelle» si elle est la conséquence directe de l'exposition d'un travailleur à un risque physique, chimique, biologique, ou résulte des conditions dans lesquelles il exerce son activité professionnelle.

Les maladies professionnelles sont définies par des tableaux, elles sont au nombre de **98**.

Chaque tableau comporte :

- ➔ les symptômes ou lésions pathologiques que doit présenter le malade,
- ➔ le délai de prise en charge : délai maximal entre l'apparition de l'affection et la date à laquelle le travailleur a cessé d'être exposé au risque,
- ➔ les travaux susceptibles de provoquer l'affection.

Parmi les maladies professionnelles on cite les principales:

- Affections péri articulaires : 65,7 %
- Affections provoquées par les poussières d'amiante : 12,32 %
- Affections chroniques du rachis lombaire dues aux charges lourdes : 7,42 %
- Affections provoquées par les bruits : 2,04 %
- chroniques du rachis lombaire dues aux vibrations : 1,58 %

IV. IMPORTANCES DES ACCIDENTS DU TRAVAIL :

1. Histoire:

Pendant la deuxième guerre mondiale, les pertes militaires enregistrées par les U.S.A. étaient, en moyenne, mensuellement de **8.126** victimes dont **3.462** tués. Pendant la même période, pour le même pays, les pertes engendrées par le travail étaient en moyenne, mensuellement, de **160.747** victimes dont **1.219** tués. Le travail faisait donc **20** fois plus de victimes que la guerre, bien que **3** fois moins de tués. Toutes les bonnes volontés sont prêtes à partir en croisade contre la guerre. Sont-elles aussi conscientes de la nécessité d'une croisade contre les accidents du travail ?

2. En 2001 (en France):

730 décès suite à un accident du travail	**635** décès suite à un accident de trajet	**318** décès suite à une maladie professionnelle

CHAQUE ANNÉE, LES ACCIDENTS DU TRAVAIL ET LES MALADIES PROFESSIONNELLES TUENT !!!

V. COUT ET CONSEQUENCES HUMAINES DES ACCIDENTS DU TRAVAIL :

1. Sur le plan national :

Si incomplètes et imparfaites qu'elles soient souvent, les données statistiques sur les accidents du tavail font apparaître l'ampleur du problème. En effet, les statistiques récentes (2001) indiquent que sur **17.233.914** salariés de l'industrie et du commerce française, inscrits au régime général de la sécurité sociale, il y a eu :

- **737.499** accidents ayant entraîné un arrêt de travail dont:
  - **43.078** accidents graves, c'est à dire ayant entraîné une incapacité permanente
  - **86.144** accidents du trajet
  - **24.220** maladies professionnelles.

La perte pour l'économie nationale se chiffre, en ne tenant compte que de l'incapacité temporaire (**Durée moyenne d'une IT = 43 jours**) à plus de **31 897 526 (≈ 32 millions)** de journées perdues dans l'année, soit l'activité totale de plus de **110.775** personnes.

2. Pour les accidentés :

- Un accidenté en état d'incapacité temporaire touche, tous les jours que dure son incapacité, une indemnité égale à la moitié ou un peu plus de son salaire journalier et cela jusqu'à la guérison ou la consolidation, c'est à dire jusqu'au jour ou l'invalidité est considérée comme définitive.
- En plus l'accidenté n'a pas à payer les frais de traitements médicaux, chirurgicaux ou pharmaceutique qui sont réglés par la caisse de sécurité sociale.
- Enfin, en cas d'invalidité permanente, une rente variable est versée en fonction du degré d'invalidité.
- Bien entendu, les ayants droit de la victime d'un accident mortel touchent également une rente.

Les indemnités et les rentes dont nous venons de parler compensent, dans une certaine mesure, la perte de ressources entraînée par l'accident de travail mais :

- Aucune rente ne peut compenser pour la veuve et les orphelins, la perte de l'époux ou du père
- Aucune rente ne peut compenser, pour la victime, la perte de la vue ou d'un membre.
- Aucune indemnité ne peut compenser la souffrance consécutive à un accident.

3. Pour l'entreprise :

Les charges résultant de la réparation des accidents de travail ne peuvent pas ne pas être incorporées aux prix de revient.

L'entreprise a à supporter :

- Le payement intégral du salaire du jour où s'est produit l'accident ce qui représente pour l'ensemble de l'industrie en France plus de **1.000.000** de journées partiellement chômées :
- La perte de temps de travail des compagnons de l'accidenté qui interrompent leur travail :
  - Par curiosité
  - Par amitié pour la victime
  - Pour secourir la victime
- La perte de temps des agents de maîtrise, les chefs de service et autres cadres :
  - Pour secourir la victime
  - Pour chercher les causes de l'accident
  - Pour remettre l'ordre de travail
  - pour trouver un autre salarié qui puisse remplacer la victime, le former ou le mettre au courant
  - pour établir le rapport d'accident ou répondre aux convocations des agents de l'autorité publique.
- Les pertes dues aux perturbations dans l'organisation des ateliers, des équipes de production
- Les frais accompagnants la mise au courant du travailleur appelé à remplacer la victime pendent son incapacité
- Les pertes dues aux détériorations des machines, d'outils, de produit ou pièces fabriquées accompagnants souvent l'accident corporel
- Les pertes dues aux retards de livraison et de la contre publicité qui les accompagne

- La perte résultant du manque de productivité de la victime et de l'équipement.

VI. PRINCIPAUX INDICATEURS DES ACCIDENTS DU TRAVAIL

Pour comparer et suivre l'évolution de l'état de sécurité, de deux entreprises, il faut tenir compte de l'effectif des travailleurs employés dans chacune d'elles. On peut y parvenir en calculant un certains nombre d'indicateurs, à savoir: le taux de fréquence des cas de lésions, l'indice de fréquence, le taux de gravité ainsi que l'indice de gravité.

1. Taux de fréquence TF:

C'est le nombre des cas de lésions par million d'heures de travail effectuées par toutes les personnes exposées au risque :

$$TF = \frac{\text{nombre total d'accidents avec arrêt}}{\text{nombre d'heures travaillées}} \text{ x } 10^6$$

2. Indice de fréquence IF:

C'est le nombre des cas de lésions par mille les personnes exposées au risque :

$$IF = \frac{\text{nombre total d'accidents avec arrêt}}{\text{nombre de salariés}} \text{ x } 10^3$$

3. Taux de gravité TG:

Après le calcul du taux et de l'indice de fréquence, qui ne considèrent que le nombre des lésions professionnelles, on calcul le taux de gravité.

$$TG = \frac{\text{nombre de journées perdues pour incapacité temporaire}}{\text{nombre d'heures travaillées}} \text{ x } 10^3$$

et enfin :

4. Indice de gravité IG:

$$IG = \frac{\text{total des taux d'incapacité permanente}}{\text{nombre d'heures travaillées}} \text{ x } 10^6$$

Exemple : Une entreprise occupe 500 personnes, qui travaillent chacune 50 semaines par an et 48 heures par semaine. Le nombre des cas de lésion professionnelle, au cours d'une année, a été de 60. Le nombre de journée chaumé en conséquence a été de 528. Pour cause de maladies ou d'accidents

et pour d'autres raisons, les travailleurs ont été absents pendant 5 pour cent du nombre total possible d'heures de travail.

Déterminer:

- Le taux de fréquence
- L'indice de fréquence
- et le taux de gravité

dans cette entreprise durant l'année considérée.

Réponse :

Nombre des heures programmées : Np = 500 x 50 x 48 = 1200.000 heures

Le nombre des heures effectuées : Ne = Np (1 – 0.05) – 528 * 8

= 1200.000 * 0.95 – 528 * 8

= 1135.776 heures

Le taux de fréquence peut alors être calculé: TF = 60 x 10^6 / 1135.776 = 52,82

L'indice de fréquence est: IF = 60 x 10^3 / 500 = 120

Le taux de gravité peut être calculé : TG = 528. 10^3 / 1135.776 = 0.46

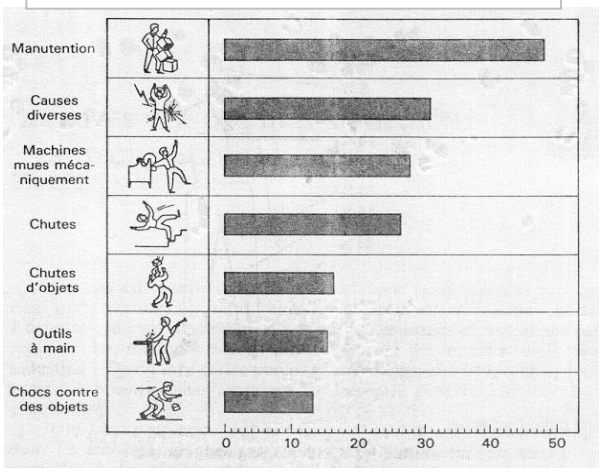
Le taux de fréquence signifie qu'au cours de l'année considérée, 53 cas de lésions professionnelles environ se sont produits par million d'heures de travail effectuées par les personnes exposées au risque.

VII. COMMENT REDUIRE LE NOMBRE DES ACCIDENTS ?

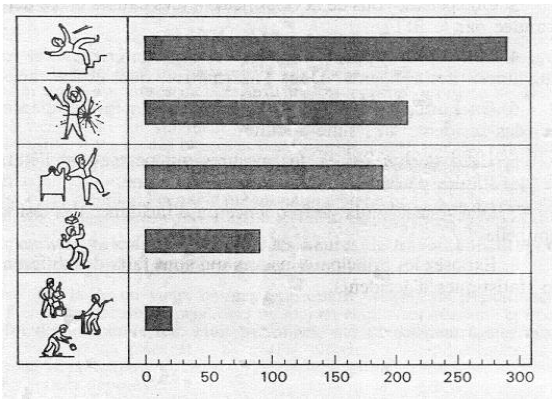
Les graves conséquences des accidents du travail amène donc à se demander s'il est possible d'en réduire le nombre. La réponse ne peut être qu'affirmative. Mais pour garantir l'obtention des résultats il faut accepter dans l'immédiat certains sacrifices qui porteront leurs fruits à plus longue échéance (durée).

Pour réduire le nombre des accidents, il faut en rechercher les causes afin de les supprimer et non de se contenter d'en rechercher la responsabilité. Ces causes varient évidemment très largement suivant les industries considérées, mais on constate que, quelle que soit l'activité exercée, les accidents les plus fréquents sont provoqués par des manutentions, des chutes de personnes ou chutes d'objets.

Nombre total d'accident par année selon la cause (en milliers)



Nombre d'accidents mortels par année selon la cause (en unités)



D'une façon sommaire, en France les principales causes d'accidents sont comme suit :

		Nombre	Décès
Manutention manuelle	**:**	**34,8 %**	**: 1,7 %**
Chute de plain pied	**:**	**21,4 %**	**: 2,7 %**
Chute de hauteur	:	**12,6 %**	**: 17,1 %**
Outils	:	6,8 %	: 0,2 %
Masse en mouvement	:	6,1 %	: 8,5 %
Machines	:	4 %	: 2,3%
Manutention mécanique	:	3,7 %	: 6,0 %
Véhicules	:	3,3 %	: 51 %
Engins TP	:	0,16 %	: 1,9 %
Électricité	:	0,11 %	: 3,11 %

On voit que, d'après cette statistique, ce n'est pas la machine qui est responsable du plus grand nombre des accidents, mais au contraire l'homme (maladresse, manque d'attention, imprudence, ignorance, négligence…).

Exemples :

- Un jeune ouvrier met en route une perceuse, mais en oubliant de retirer la clé du mandrin porte-foret. La clé est projetée au loin, blesse l'ouvrier ou un de ses compagnons.

L'accident est bien survenu à l'occasion de l'utilisation d'une machine, mais est dû en réalité à la négligence ou à l'ignorance.

- Un ouvrier, pour voir si une ligne 110 V est sous tension, la touche du doigt. Par malheur, son autre main est en contact avec tuyauterie métallique constituant un excellent conducteur de retour à la terre. L'ouvrier est électrocuté.

L'accident est bien dû à l'électricité, mais il a été provoqué par une imprudence.

- Un mécanicien, pour réparer un tour parallèle, enlève le protecteur des engrenages de la tête de cheval. La réparation terminée, il omet de remettre le protecteur en place. Pendant qu'il essaie la machine, un de ses compagnons passe à proximité du tour et se fait happer le doigt.

Là aussi, l'accident est le résultat d'une négligence grave.

1. Mesure d'ordre technique :

D'une façon générale, la prévention des accidents doit commencer par l'étude des risques.

a. Lors de l'établissement d'un projet de bâtiment, d'unité, d'appareil, de machine, etc. , tenir compte de toutes les exigences de la sécurité (et non seulement en ce qui concerne les dangers graves).

b. Maintenir tous les appareillages, machines, bâtiments, etc. en parfait état d'entretien (l'ordre et la propreté, facteurs importants de la sécurité, font partie du bon entretien).

c. Vérifier très fréquemment :
   - Le petit outillage
   - Les appareils de levage et de manutention
   - Les installations électriques
   - Les installations de compression de fluides divers
   - Les installations vapeur
   - Les échelles mobiles, etc.
   - Tous les appareils de contrôle et de sécurité (manomètres, thermomètres, soupapes de sécurité, etc.).

d. Veiller à l'utilisation rationnelle des équipements de protection :
   - Lunettes individuelles
   - Chaussures spéciales
   - Gants
   - Tabliers, combinaisons spéciales

- Masques, écrans spéciaux
- Masques respiratoires
- Ceintures de sécurité pour le travail sur échafaudage, etc.

e. Assurer le respect de certaines consignes de travail qui doivent être absolument impératives :
   - Défense de fumer sauf dans les emplacements spécialement autorisés
   - Défense de nettoyer une machine en marche
   - Défense de remonter les courroies en marche
   - Défense d'utiliser une machine dont les protecteurs ne sont en place
   - Défense de travailler sur les lignes électriques sous tension
   - Défense de resserrer des joints sur des appareils ou tuyauteries sous pression, etc.

f. Chaque fois qu'un travail fait l'objet d'un planning, en même temps les conditions de sécurité de travail seront étudiées.

g. Les manutentions sont, à l'origine d'une proportion très importante des accidents, donc il est judicieux chaque fois que cela est possible de remplacer les manutentions manuelles par des moyens mécaniques.

2. Sélection médicale et psychotechnique du personnel :

Enfin, et ce n'est pas le moins important des facteurs de réduction du nombre des accidents, il faut ne confier à chaque travailleur que des tâches que ses aptitudes, tant physiques que psychiques, lui permettent de remplir sans danger pour lui et pour ses camarades.

Il n'est pas besoin de dire qu'il ne faut pas confier des travaux de manutention lourde à un gringalet (petit homme maigre), mais il ne faut pas non plus les confier à un homme d'apparence robuste mais ayant une faiblesse de la paroi inguinale, ou ayant une maladie de cœur, etc.

On ne devra pas confier la conduite d'un pont roulant, d'une grue, à un ouvrier ayant une inaptitude innée à apprécier correctement les distances.

C'est dire assez l'importance de la sélection médicale et psychotechnique qui est absolument nécessaire pour tous les travaux mettant en jeu une source d'énergie.

3. Mesure d'ordre psychologique :

- Un protecteur (couvercle, grillage …) parfaitement bien conçu, c'est très bien, à condition que le compagnon ne le démonte pas pour travailler plus à l'aise.

- Un petit outillage en parfait état, c'est aussi très bien, à condition que l'ouvrier n'utilise par une clé de 18 pour serrer un écrou de 12 en interposant une cale mobile.

Ceci pour rappeler que tous les dispositifs, aussi bon qu'ils soient, toutes les consignes, aussi bien étudiées qu'elles puissent être, n'auront que bien peu d'efficacité, ou même pas du tout, si les ouvriers ne sont pas prudents, réfléchis, attentifs, en un mot : s'ils n'ont pas **l'esprit de sécurité**.

Pour développer l'esprit de sécurité, de nombreux moyens peuvent être mis en œuvre :

a. **Les affiches illustrées :** affiches montrant comment les accidents se produisent et ce qu'il faut faire pour les éviter, ou mieux, affiches montrant la façon non dangereuse d'exécuter une opération.

   Le choix de l'emplacement des affiches doit être judicieux. Ne pas oublier que les premiers jours suivant l'apposition d'une affiche on la regarde, après, elle fait partie du décor, on ne la voit plus. Il faut donc renouveler fréquemment le sujet affiché, ou varier la présentation du même sujet.

b. **Les panneaux d'affichage des résultats :** Sont surtout utiles dans les établissements à effectifs importants et peuvent servir à développer l'émulation entre les équipes. Ils doivent être d 'une lecture facile et indiquer par exemple : le nombre d'accidents depuis le début du mois, le nombre de jours écoulés depuis le dernier accident, le record à battre.

   Pour être utiles, il faut que ces tableaux soient absolument sincères et tenus scrupuleusement à jour.

c. Les articles dans le journal d'entreprise

d. **Les films :** Ils constituent un excellent moyen de propagande. Ils coûtent fort cher et son très difficiles à bien réaliser.

e. **Les causeries de sécurité :** Elles constituent un des meilleurs moyens pour développer l'esprit de sécurité.

   Ces causeries qui peuvent durer un quart d'heure se feront sur le lieu même du travail, soit au début, soit à la fin de la journée. On y passera en revue les accidents récemment survenus dans l'atelier ou dans les ateliers voisins et on demandera aux auditeurs eux-mêmes d'indiquer les moyens d'en éviter le retour

   Ces causeries devront être faites par des personnes connaissant parfaitement le travail effectué par l'équipe.

EXEMPLES DE TAUX DE FRÉQUENCE

par comités techniques nationaux

Bâtiment et travaux public	: 57,6
Bois, ameublement, papier carton	: 35,1
Services, commerces et industrie de l'alimentation	: 33,3
Métallurgie	: 27,9
Transport, eau, gaz, électricité	: 25,2
Chimie, caoutchouc, plasturgie	: 23,2
Commerce non alimentaire	: 14,5
Activités de service et travail temporaire	: 27,6

Moyenne 24,6

EXEMPLES DE TAUX DE GRAVITÉ

par comités techniques nationaux

Bâtiment et travaux public	: 2,95
Bois, ameublement, papier carton	: 1,45
Services, commerces et industrie de l'alimentation	: 1,29
Métallurgie	: 1,01
Transport, eau, gaz, électricité	: 1,22
Chimie, caoutchouc, plasturgie	: 0,88
Commerce non alimentaire	: 0,65
Activités de service et travail temporaire	: 1,22

Moyenne 1,06

EXEMPLES D'INDICE DE GRAVITÉ

par comités techniques nationaux

Bâtiment et travaux public	: 49,1
Bois, ameublement, papier carton	: 21,4
Services, commerces et industrie de l'alimentation	: 12,8
Métallurgie	: 15,5
Transport, eau, gaz, électricité	: 18,0
Chimie, caoutchouc, plasturgie	: 13,2
Commerce non alimentaire	: 9
Activités de service et travail temporaire	: 12,4

Moyenne 14,5

Document pédagogique N°2 :

TIME UNIVERSITE

Cours Développement des Logiciels

3ème année GL

Melle Inès AMMARY

Année Universitaire 2008-2009

Plan du cours Développement de logiciels

Chapitre 1 : Du Système d'Information au Logiciel

1. Définition d'un Système d'information d'une organisation

2. Définition d'un Système d'information automatisé

3. Les SI face au progrès technologique

4. Informatisation des SI

    4 .1 Définition du génie logiciel (Software Engineering)

    4 .2 Objectifs du génie logiciel

Chapitre 2 : Processus de développement : Principaux modèles

1. Le rôle des modèles

2. Modèles de cycles de vie linéaires

    2.1 Définitions d'un cycle de vie

    2.2 Durée d'un cycle de vie

    2.3 Grandes phases d'un cycle de vie

3. Le modèle en cascade (chute d'eau)

    3.1 Phase d'étude préalable (ou d'opportunité)

    3.2 Phase de spécification

    3.3 Phase de spécification fonctionnelle

    3.4 Phase de conception préliminaire (générale)

    3.5 Phase de conception détaillée

    3.6 Phase de codage

    3.7 Phase de tests unitaires

Chapitre 1 : Du Système d'Information au Logiciel

1. Définition d'un Système d'information d'une organisation :

Toute organisation peut être considérée comme un système traitant des flux physiques et des flux d'informations.

➔ Un système d'information (SI) peut être considéré comme un ensemble de flux d'informations, d'opérations qu'ils subissent et de moyens mis en œuvre pour ce faire quelque soit la nature de ces moyens.

➔ Un système d'information est défini par l'ensemble de moyens humains, matériels et méthodes se rapportant au traitement des différentes formes d'informations rencontrées dans les organisations.

➔ Un système d'information peut être constitué de procédures manuelles ou automatisées.

2. Définition d'un Système d'information automatisé :

➔ Un système d'information automatisé (SIA) est l'ensemble des moyens et des méthodes se rapportant au traitement automatisé des données de l'organisation.

➔ Les SIA sont perçus à travers les logiciels qui les composent.

➔ Les SIA d'une organisation ne forment pas un tout homogène mais un ensemble de logiciels qui sont :

- élaborés à des dates différentes,
- dans des environnements informatiques différents,
- partageant certaines ressources (bases de données, matériel, …)

3. Les SI face au progrès technologique :

- Accroissement de la puissance du matériel (espace de stockage, vitesse du micro processeur, …)
- Abaissement du coût du matériel :
  - Chute du coût relatif des composants matériels d'un système informatique.
  - Croissance du coût relatif des composants logiciels.
- Diversification du champ d'application de l'informatique.
- Demande de logiciels sans cesse croissante.
- Réalisation de tâches de plus en plus complexes.

4. Informatisation des SI :

➔ Les démarches d'informatisation sont aujourd'hui connues et bien définies, mais il subsiste encore de nombreux problèmes d'ordre à la fois économiques, politiques, sociaux et organisationnels.

➔ De nombreux projets échouent en raison du non respect du cahier des charges à la fois en ce qui concerne les budgets et les délais.

➔ Les logiciels livrés ne sont pas toujours de qualité c'est à dire ils ne répondent pas aux besoins des utilisateurs.

➔ Aussi parle-t-on toujours de crise de logiciel ?

Le processus d'informatisation des SI comprend deux activités principales :

- Activité de développement,
- Activité de maintenance.

➔ Ces activités sont mises en œuvre suivant des stratégies diverses et parfois complexes requérant des compétences variées dans de nombreux domaines de l'informatique à savoir les bases de données, les techniques et les outils de programmation, les réseaux … Ce qui a donné naissance au Génie logiciel.

4 .1 Définition du génie logiciel (Software Engineering) :

Le génie logiciel est un domaine relativement récent :

- La première conférence sur le thème du génie logiciel eut lieu en 1968 (conférence de l'OTAN).
- *Génie logiciel* : élaboration et l'utilisation des principes de génie permettant de produire économiquement des logiciels fiables et qui fonctionnent de façon efficace sur des machines réelles.
- *Génie logiciel* : application pratique de la connaissance scientifique dans la conception et l'élaboration de programmes informatiques et de la documentation associée nécessaire pour les développer, les mettre en œuvre et les maintenir (Bohem 1976).
- *Génie logiciel* : ensemble des activités de conception et de mise en œuvre des produits et des procédures tendant à rationnaliser la production du logiciel et son suivi.
- Le génie logiciel consiste à appliquer des méthodologies*1 c'est-à-dire à développer et à utiliser des méthodes*2 et des outils*3 dans le but de produire un logiciel de qualité en respectant les contraintes de temps et de coûts.

*1 Méthodologie : c'est un ensemble structuré et cohérent de méthodes et d'outils permettant de déduire la manière de résoudre un problème. Exemples de méthodologies : MERISE (Méthode d'Etude et de Réalisation informatiques des Systèmes d'Entreprise) et le Processus Unifié.

*2 Méthode : c'est un ensemble de concepts, de techniques d'aide à la résolution d'un problème. Elle consiste en une démarche rationnelle de l'esprit pour arriver à la connaissance ou à la démonstration de la vérité. Exemples de méthodes : MERISE et UML (Unified Modelling Language).

*3 : c'est un ensemble de programmes, de langages et de formulaires de documentation qui peuvent aider à la mise en œuvre des techniques. Exemples d'outils : AMC Designer et Ratioanal Rose.

Dire qu'un logiciel est de qualité sous entend qu'on puisse lui appliquer certains critères comme :

- L'adéquation aux besoins des utilisateurs,
- La fiabilité,
- L'efficacité,
- L'évolutivité.
- …

4.2 Les objectifs du génie logiciel

L'objectif primordial du génie logiciel consiste en le développement et la production du logiciel et son suivi.

C'est quoi un logiciel ?

Un logiciel n'est pas que du code !

C'est un ensemble de :

- Programmes,
- Procédés,
- Règles,
- Documentation.

Un logiciel est caractérisé par :

- Une diversité des applications : un logiciel est un produit atypique comparativement aux produits industriels : on peut parler d'avion type, de voiture type mais il est quasiment impossible de parler de logiciel type.
- Une taille et une complexité
- Abstraction et invisibilité
- Coût élevé du logiciel : les logiciels coûtent cher du fait leur taille et leur complexité inhérente. La nature des logiciels rend difficile une estimation à priori des coûts et des délais. Souvent, on a coutume de prévoir une marge chronologique et financière de 30% en plus.
- Evolutivité des besoins : un logiciel est un produit continuellement évolutif étant destiné à satisfaire un ensemble de besoins appartenant à un monde réel qui évolue. Ces besoins

évoluent et le logiciel que l'on développe est amené à évoluer sinon il sera replacé par un autre logiciel répondant mieux à ces nouveaux besoins.

- Distribution et parallélisme :
  - De plus en plus de parallélisme :

Le traitement étendu des logiciels à des activités naturellement parallèles amène à la conception et à la réalisation de composants (modules logiciels) parallèles entre les quels il peut y avoir : communication de données ou de messages, synchronisation, conflits ou concurrence pour utiliser une même ressource.

  - De plus en plus de distribution :

C'est le cas des applications où les données et les traitements sont répartis sur plusieurs sites plus ou moins distants.

4 .2 Objectifs du génie logiciel :

Le génie logiciel a comme objectif l'utilisation de l'informatique dans les entreprises comme outil de tous les jours :

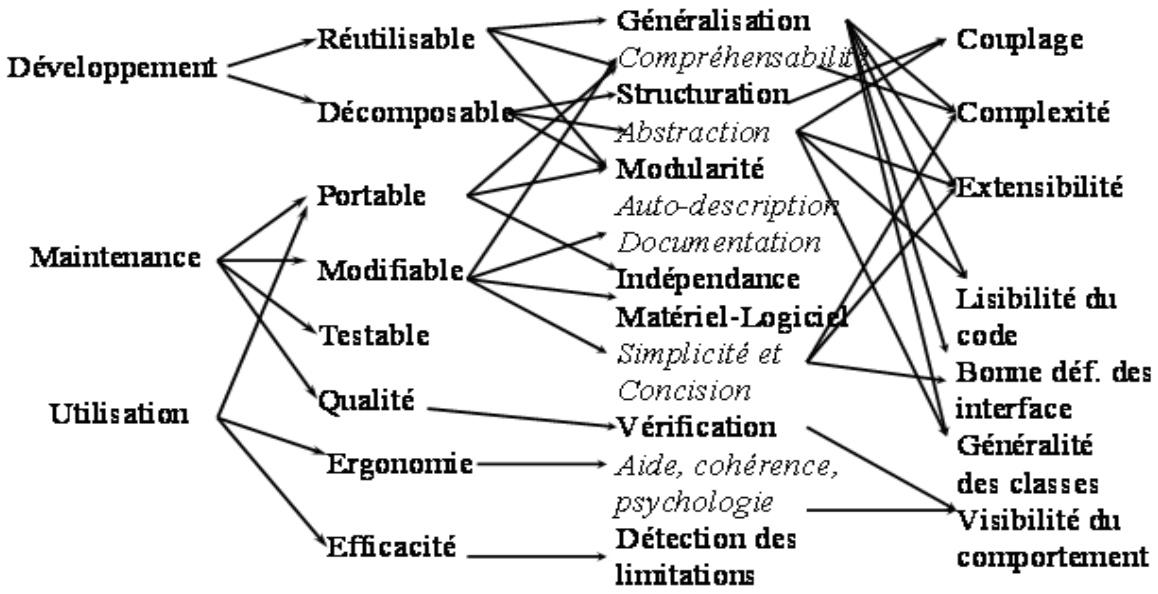
- Besoin d'acquérir et produire des logiciels,
- Passer à une production industrielle des logiciels de telle sorte qu'ils soient fiables et peu coûteux (réduction des coûts de production, de maintenance et d'utilisation.

L'objectif étant donc de maîtriser le développement c'est-à-dire réduire les coûts et assurer la qualité grâce à l'utilisation :

- de méthodes (façon de faire quelque chose,
- d'outils (programmes, langages, ..)
- de méthodologies.

Les objectifs fixés ne peuvent être atteints que si certains principes sont suivis. Chaque méthodologie privilégie certains principes et le choix d'une méthodologie pour un projet va dépendre des objectifs visés.
Le choix sera forcément un compromis puisque aucune méthodologie ne pourra couvrir tous les principes devant être suivis.
L'évaluation de la qualité du logiciel se fera par la mesure des attributs de chacun de ces principes.

Coûts Objectifs Principes Attributs



❖ Objectifs :

- *Réutilisable* : possibilité d'utiliser certaines parties du code pour résoudre un autre problème.
- *Décomposable* :
- *Portable* : facilité avec laquelle un logiciel peut être transféré sous différents environnements matériels et logiciels.
- *Ergonomie* : l'ergonomie logicielle a pour objectif l'amélioration de l'interaction Homme-Ordinateur ; faire en sorte que toute application informatique livrée aux utilisateurs soit :

  ➔ Utile : l'outil réalisé doit répondre aux besoins des utilisateurs pour lesquels il a été conçu, autrement dit qui soit en adéquation avec leur tâche.

  ➔ Utilisable (maniable) : c'est-à-dire facile à utiliser.

  L'ergonomie d'une application se situe à deux niveaux :

  ➔ Ergonomie de surface : ce que voit l'utilisateur de l'interface : structuration des éléments dans les fenêtres / pages (dispositions de ces éléments, densité d'affichage, ..), aspects graphiques.

  ➔ Ergonomie profonde : fonctionnalités offertes à l'utilisateur, structure de l'outil autrement dit la répartition des fonctionnalités en modules, menus, …, les enchaînements d'écrans qui traduisent la succession des tâches que l'utilisateur a à effectuer.
- *Efficacité* : en termes de rapidité d'exécution et taille mémoire, utiliser de manière optimale les ressources matérielles.

❖ Principes :

- *Généralisation* : regroupement d'un ensemble de fonctionnalités semblables en une fonctionnalité paramétrable (généricité, héritage).
- *Structuration* : façon de décomposer un logiciel : utilisation d'une méthode Botton-Up ou Top-Down ;
- *Abstraction* : mécanisme qui permet de présenter un contexte en exprimant les éléments pertinents et en omettant ceux qui ne le sont pas.
- *Modularité* : décomposition d'un logiciel en des composants discrets.
- *Documentation* : gestion des documents incluant leur identification, acquisition, production, stockage et distribution.
- *Vérification* : détermination du respect des spécifications établies sur la base des besoins identifiés dans la phase précédente du cycle de vie.

❖ Attributs :

- *Couplage* : mesure des interconnexions entre modules.
- *Complexité* : degré de compilation d'un système.
- *Extensibilité* : degré d'accommodations aux changements produits par une nouvelle exigence à satisfaire.
- *Lisibilité* : difficulté de comprendre un composant logiciel (lié à la complexité et à la documentation).
- *Bonne définition des interfaces* : mesure du degré de dépendance fonctionnelle entre les tâches réalisées par un module.
- *Généralité des classes ou degré d'abstraction* : une classe est caractérisée par les opérations qu'elle permet d'effectuer.
- *Visibilité du comportement* : représente la possibilité de se rendre compte de façon aisée de ce que font les différentes fonctionnalités d'un programme.

Série 1 : Du Système d'Information au Logiciel

Questions

1. Faites un rapprochement entre Merise et UML ;

2. Faites un rapprochement entre Merise et le Génie Logiciel.

3. Faites le rapprochement entre UML et le Génie Logiciel.

– Définissez les termes méthode, méthodologie et outils.

Correction Série 1 : Du Système d'Information au Logiciel

Question 1

- *Présentation*

Merise	UML
Franceest une méthode de conception, de développement et de réalisation de projet informatique. Cette méthode Merise est basée sur la séparation Données-Traitements, cette séparation assure une longévité au modèle. En effet, les données sont relativement stables alors que les traitements peuvent évoluer. La vocation de Merise est double : - Représente une méthode de conception de systèmes d'informations. - Représente une méthode méthodologique de développement de systèmes informatiques. *Démarche :* Découpage du Processus de développement en étapes et phases de type Cascade.	C'est un langage de modélisation objet. La version 1.1 a été remise à l'OMG et a été approuvée comme standard en 1997. ▪ Il facilite l'expression et la communication de modèles en fournissant un ensemble de symboles. ▪ Il fournit des fondements pour : - Spécifier : modéliser le *Quoi Faire* et le *Comment Faire*. - Visualiser : les modèles construits seront de base à la réalisation informatique. - Documenter : les modèles sont utilisés pour communiquer les connaissances. ▪ Il définit 9 diagrammes : - diagramme d'objets - Diagramme de classes - Diagramme d'état transition - Diagramme de cas d'utilisation - Diagramme de séquence - Diagramme de collaboration - Diagramme d'activités - Diagramme de composants - Diagramme de déploiement *Démarche :* Itérative et incrémentale.

- *Points communs*

Merise	UML
Entité Relation Attribut Cardinalité Occurrence Généralisation Spécification Acteur Flux	Classe Association Attribut Multiplicité Objet (Instance) Généralisation Spécification Acteur Message

- *Concepts spécifiques*

Merise	UML
MCD MOD MLD MCC . .	L'encapsulation Le polymorphisme Agrégation Diagramme de cas d'utilisation Diagramme de collaboration Diagramme de séquence Diagramme de déploiement

Question 2-3

Le Génie Logiciel est une discipline de l'informatique qui s'intéresse au développement, maintenance, utilisation, outils, méthodes et activités de gestion. Il consiste en une pratique de la connaissance scientifique dans la conception et l'élaboration de programmes informatiques et de la documentation associée nécessaire pour les développer, les mettre en œuvre et les maintenir [Boehm 1976].

Merise représente une méthode de conception, développement et réalisation de projets informatiques. Ainsi donc, elle peut être conçue comme une des méthodes utilisées en génie logiciel décrivant un processus discipliné qui génère un ensemble de modèles décrivant les différents aspects d'un système logiciel, en utilisant une certaine notation bien définie.

UML est un langage de modélisation favorisant la communication entre le développeur et le gestionnaire à travers un ensemble de diagrammes modélisant le système à développer.

Ainsi donc, UML est considéré comme un langage de modélisation et par conséquent un outil utilisé en génie logiciel.

Question 4

- *Définition d'un système d'information* :C'est la partie du réel constituée d'informations organisées, d'évènements ayant un effet sur ces informations et d'acteurs qui agissent sur ces informations. Le système d'information intègre les dimensions organisationnelle, humaine et technologique de la gestion de l'information de l'entreprise.

- *Définition d'un système informatique* : C'est un ensemble organisé d'objets techniques (matériels, logiciels de base, application) dont la mise en œuvre réalise l'infrastructure d'un système d'information.

- *Définition d'un système d'information automatisé* : C'est une partie du système informatique regroupant uniquement les applications.

sur une version opérationnelle du logiciel, par contre les premiers cycles, ils servent essentiellement à clarifier les besoins.

- *Avantages du modèle en spirale*

➔ Le caractère itératif du modèle,

➔ La considération des risques pouvant être appréhendés dès le début du projet permettant ainsi de réviser les choix effectués au cours des différents cycles en fonction de l'avancement du projet,

➔ La rapidité de production d'un logiciel opérationnel même s'il est minimale,

➔ Il permet de se concentrer sur les aspects les plus incertains du développement,

➔ Il tolère la remise en cause de la part du client à chaque nouvelle évaluation

- *Inconvénients du modèle en spirale*

➔ Risque de remise en cause des spécifications des versions déjà réalisées lors de l'analyse de nouvelles versions,

➔Difficultés de mise en œuvre au niveau procédural et de contrôle du processus,

➔ Organisation opérationnelle du développement souvent modifiée pour le client,

➔ Difficultés pour mener à bien les premiers cycles de la spirale.

Document pédagogique N°3 :

Questy

L'évaluation des connaissances

L'élaboration des questions à choix multiples

Dominique Bonnefon

Les distracteurs doivent être crédibles et pertinents :
Il est tentant d'introduire une note d'humour sous la forme d'un distracteur fantaisie, mais il faut savoir que dans ce cas, on incite les élèves à procéder par élimination au lieu de raisonner selon le schéma classique prévu par le concepteur.

Fréquemment, l'élève peut éliminer rapidement une ou deux réponses puis il hésite entre deux autres choix …

Il convient d'éviter que la bonne réponse soit systématiquement plus longue que les autres ou qu' elle soit placée au milieu de la liste des autres propositions

Les questions à choix multiples peuvent se présenter sous différentes formes

1 – Réponse binaire

Une affirmation est faite, deux réponses sont proposées : soit **Vrai / Faux** , **Oui / Non** ou **d'accord / Pas d'accord :**

- **La lune est un satellite de la Terre :**

❑ Vrai
❑ Faux

2 – Réponse unique

une affirmation est énoncée, plusieurs réponses sont proposées, une seule est valide :

- **la capitale de la France est :**

❑ Lyon
❑ Paris
❑ Versailles
❑ Vichy

3 – Réponses multiples

Plusieurs réponses sont proposées, la bonne réponse exige de cocher plusieurs cases :

- **Quelles sont les villes capitale d'état ?**

- ❑ Londres
- ❑ Paris
- ❑ Rome
- ❑ Barcelone

Pour obtenir une bonne réponse, le candidat **doit** cocher toute les bonnes réponses ; un seul oubli rend sa réponse fausse !

4 – Réponses en énumération classée

Lorsque la réponse exacte comporte plusieurs éléments, il est possible d'utiliser la méthode suivante :

- **Les couleurs du drapeau du Royaume de Belgique sont :**

- ❑ Bleu, blanc, rouge
- ❑ Rouge, blanc, jaune
- ❑ Noir, jaune, rouge
- ❑ Rouge, vert, noir

5 – Réponses en énumération classée (variante)

Cette variante de l'énumération classée, demande plus d'attention que la question précédente ; elle consiste à proposer les éléments sous forme de liste numérotée dans le corps de la question (cela permet de dépasser le nombre de 5 propositions) :

- **Les couleurs du drapeau du Royaume de Belgique sont :**
    1. **Rouge**
    2. **Vert**
    3. **Jaune**
    4. **Noir**
    5. **Bleu**
    6. **Orange**

- ❑ 1, 2 et 3
- ❑ 3, 4 et 1
- ❑ 4, 3 et 1
- ❑ 6, 2 et 3

6 – Réponses multiples équivalentes

Si deux réponses sont possibles, il convient de déterminer si une seule réponse suffit à valider la réponse ou pas :

- L'unité astronomique est égale à :
- ❑ la distance de la terre au soleil
- ❑ la distance du soleil à l'étoile la plus proche
- ❑ 149,6 millions de km
- ❑ 384 400 km
- ❑ la distance de la terre à la lune

commentaire : les réponses « la distance de la terre au soleil » et « 149,6 millions de km » sont exactes ; si l'élève indique l'une ou l'autre réponse, il est logique de considérer que sa réponse est correcte ; le commentaire qui suit l'annonce du résultat, peut apporter des précisions sur la dualité de réponse. Si le réalisateur de questionnaires estime que deux réponses doivent être apportées pour valider la réponse, il devra rédiger sa question différemment, en employant la méthode précédente ; exemple :

- L'unité astronomique est égale à :
- ❑ la distance de la terre au soleil, soit 149,6 millions de km
- ❑ la distance de la terre à la lune soit 384 400 km
- ❑ la distance de la terre au soleil soit 2,4 milliards de km
- ❑ la distance de la terre à la lune soit 98,5 millions de km

Commentaire : dans ce cas, seule la première réponse est considérée comme exacte. Comme souvent, le commentaire pourra expliquer les erreur que représente les autres propositions (la distance terre-lune est bien de 384 400 km en moyenne, mais ce n'est pas l'unité astronomique…)

Une autre solution possible

- L'unité astronomique est égale à :

1. **La distance de la terre au soleil**
2. **La distance de la terre à la lune**
3. **384 400 km**
4. **2,4 milliards de km**
5. **98,5 millions de km**
6. **149,6 millions de km**

- ❑ 1 et 3
- ❑ 1 et 4
- ❑ 1 et 6
- ❑ 2 et 3
- ❑ 2 et 5

Commentaire : ici encore, la seule réponse exacte est « 1 et 6 »

7 – Réponse par association

En règle générale, chaque religion possède un (ou des) livre(s) sacré(s) ;
Indiquez les associations correctes :

A Veda	**1 Christianisme**
B Upanishad	**2 Islam**
C Bible	**3 Hindouisme**
D Coran	**4 Shintoïsme**

- ❑ A 3
- ❑ B 4
- ❑ C 2
- ❑ D 2
- ❑ C 1

Commentaire : Cette question comporte un piège : Veda et Upanishad sont les livres sacrés de l'Hindouisme ; Le shintoïsme n'est donc pas concerné ! Les bonnes réponses étaient donc : A3, D2 et C1 …

8 – Réponse par exclusion (« chassez l'intrus »)

La question à choix multiples peut également proposer une exclusion :

- Parmi ces animaux marins, chassez l'intrus :

- ❑ le requin
- ❑ l'espadon
- ❑ le dauphin
- ❑ la morue
- ❑ le mérou

Commentaire : Le dauphin n'est pas un poisson, mais un mammifère marin. Ce type de question demande également plus d'attention, notamment si elle est glissée au milieu de nombreuses questions « positives »

9 – Question à trou

Le texte de la question se présente sous la forme d'une phrase au sein de laquelle il manque un mot (et un seul !) ; l'une des propositions est le mot manquant.

- La cigale ayant chanté tout l'été, se trouva fort dépourvue quand la *** fut venue :**

- ❑ tourmente
- ❑ brise
- ❑ saison
- ❑ bise
- ❑ neige

Document pédagogique N°4 :

Bases de données relationnelles et contraintes SQL

par Frédéric Brouard, alias SQLpro
MVP SQL Server
Expert langage SQL, SGBDR, modélisation de données

Auteur de :

- SQLpro http://sqlpro.developpez.com/
- "SQL", coll. Synthex, avec C. Soutou, Pearson Education 2005
- "SQL" coll. Développement, Campus Press 2001

Enseignant aux Arts & Métiers et à l'ISEN Toulon

L'une des idées force de la conception des bases de données relationnelles repose sur la notion de contrainte. Une contrainte n'est autre qu'une règle impérative ne devant en aucun cas être violée. Certaines contraintes sont le reflet du modèle de données et permettent d'assurer la cohérence fonctionnelle des relations entre les tables. D'autres assurent que les données saisies correspondent bien aux limites de l'univers que l'on modélise. Enfin, les règles "métiers", c'est à dire le fonctionnel applicatif, nécessite la mise en œuvre de contraintes complémentaires souvent complexes. Mais, si toutes les contraintes s'avèrent nécessaires, elles sont souvent mal comprise, très souvent mal gérées dans les processus de développement, et se voient donc souvent reléguées, voire abandonnées.
Cet article à pour but de vous montrer l'intérêt des contraintes SQL avec des exemples concret de leur utilité.

Mes audits m'ont souvent montrés que les contraintes étaient souvent peu utilisées. Je crois avoir une explication... Le mot *contrainte* fait peur. A l'origine, dans le droit, on trouve la *contrainte par corps*, une mesure d'exécution de peine qui permet l'incarcération de la personne qui ne s'acquitte pas d'une condamnation pécuniaire. Abolie en France depuis quelques décennies, elle reste en vigueur dans certains pays comme la Tunisie : *la contrainte par corps est exécutée à raison d'un jour d'emprisonnement par trois dinars ou fraction de trois dinars sans que sa durée puisse excéder deux ans* (art. 344 de la Loi n°99-90 du 2 août 1999).
Dans d'autres domaines, la notion de contrainte est quand même plus réjouissante. Il y a un peu plus d'un an, devant donner une formation sur SQL, je voyais une jeune fille, arrivée fort tôt au cours, patienter en remplissant une série de grille que je pris pour un carré magique. "Tenez lui dis-je, pourquoi ne pas résoudre votre problème en utilisant une simple requête SQL ?" Je lui promis de terminer la formation par une telle démonstration. La résolution des sudokus reposent sur l'utilisation de contraintes.
Il existe toute une branche des mathématiques et plus précisément dans l'algorithmique et dans l'informatique qui utilise massivement les contraintes. Je veut parler de la programmation par contraintes, notamment à travers un langage comme PROLOG ou encore un framework Java spécialisé et non commercial comme CHOCO.
C'est d'ailleurs à l'aide de la programmation par contrainte qu'en 1976 Appel et Haken, à l'université de l'Illinois, firent la démonstration finale du théorème de la carte à 4 couleurs, en explorant systématiquement les cas particuliers. C'était aussi la première fois que la démonstration d'un théorème se faisait à l'aide d'un ordinateur.

Ce que je vous propose de visiter c'est comment SQL implémente les contraintes dans les bases de données relationnelles.

1 – La portée des contraintes.

Dans une application qui utilise SQL, on trouve les éléments suivants : des bases de données relationnelles dans lesquelles se trouvent des tables et des vues. Tables et vues sont dotées de colonnes et les données sont écrites lignes par lignes. Finalement l'élément le plus petit de cet ensemble est la donnée. Les contraintes se trouvent à chacun des niveaux de cet édifice.
Les contraintes dites "de domaine" concernent les valeurs que revêtent les données des colonnes. Les contraintes de tables peuvent porter sur une colonne, sur une ligne ou sur une table, mais valident la cohérence de la ligne. Enfin les assertions peuvent porter sur plusieurs tables, voire toutes les tables de la base et assurent une cohérence transverse.

1.1 - Portée des contraintes de table

Les contraintes de table sont les plus connues. La plus classique ne concerne qu'une colonne à la fois. C'est la contrainte d'**obligation de valeur** (NOT NULL) qui exige, pour la colonne qui en est pourvue, qu'à toute ligne de la table une valeur soit exprimée.
On trouve ensuite la contrainte de **clef primaire** (PRIMARY KEY) qui assure l'unicité de la référence à une ligne d'une table. C'est le moyen par lequel on repère une ligne et une seule dans la table. Toutes les colonnes concourant à la clef se doivent d'être valuées (NOT NULL).
La contrainte d'**unicité** (UNIQUE) permet de s'assurer qu'une autre clef pourrait remplacer la clef primaire. Mais à la différence de la clef primaire, la contrainte d'unicité n'oblige pas à ce que les données participant à la formation de la contrainte soient valuées. Il peut même y avoir plusieurs lignes de la table dont les données formant la contrainte d'unicité sont vides de toutes valeurs. En d'autres termes, la contrainte d'unicité exige que toute donnée *valuée* soit distincte... Par essence une donnée non valuée ne peut jamais être comparée à une autre données non valuée, pas même à elle même. C'est à dire que le prédicat " NULL = NULL " ne sera ni vrai ni faux mais tout simplement inévaluable !
La contrainte la plus draconienne est la contrainte de **validation** (CHECK). Elle permet de restreindre les valeurs de la ou les colonnes qui la composent afin de respecter des règles même les plus complexes qui soient. Nous lui consacrerons un long paragraphe.

Enfin, la contrainte la plus redoutée par les développeurs, parce que mal appréhendée, est la contrainte de **clef étrangère** (FOREIGN KEY) destinée à assurer l'intégrité de référence. Nous la détaillerons.

Voici quelques exemples de ces contraintes résumé dans une table :

```
CREATE TABLE T_PATIENT_PTN
(PTN_ID           INT NOT NULL PRIMARY KEY,
 PTN_NUM_SECU     CHAR(13) UNIQUE,
 PTN_CLEF_SECU    CHAR(2)
                  CHECK (PTN_CLEF_SECU IS NULL
                     OR (SUBSTRING(PTN_CLEF_SECU FROM 1 FOR 1)
                                 BETWEEN 0 AND 9) AND
                         SUBSTRING(PTN_CLEF_SECU FROM 2 FOR 1)
                                 BETWEEN 0 AND 9)),
 PTN_NOM          CHAR(32) NOT NULL,
 PTN_PRENOM       VARCHAR(25),
 PTN_DATE_NAIS    DATE,
 PTN_CIVILITE     INT FOREIGN KEY REFERENCES T_CIVILITE_CVT ( CVT_ID))
```

Notez que le numéro de sécurité sociale est unique mais n'a pas l'obligation d'être valué. C'est pratique si vous êtes médecin et que votre patient à oublié sa carte vitale ! Pour la clef de contrôle du numéro de sécurité sociale, nous avons exigée qu'elle soit composée de deux caractères dont les valeurs peuvent être saisies entre '0' et '9'. Enfin, le code civilité doit être choisit parmi les valeurs de la clef de la table T_CIVILITE_CVT.

Les contraintes PRIMARY KEY, UNIQUE, FOREIGN KEY et CHECK peuvent être spécifiées directement dans la définition de la ligne de la table si elles ne portent que sur une seule colonne, sinon elles doivent être spécifiées en tant qu'attribut de la table. Ce dernier mode est à préférer, même pour des contraintes mono colonnes. En effet il présente plusieurs avantages, car dans ce cas la contrainte doit être nommée : son nom fera partie du message d'erreur et il sera plus facile de supprimer ou désactiver cette contrainte.

Exemple :

```
CREATE TABLE T_PATIENT_PTN
(PTN_ID           INT NOT NULL,
 PTN_NUM_SECU     CHAR(13),
 PTN_CLEF_SECU    CHAR(2),
 PTN_NOM          CHAR(32) NOT NULL,
 PTN_PRENOM       VARCHAR(25),
 PTN_DATE_NAIS    DATE,
 PTN_CIVILITE     INT,
 CONSTRAINT       PK_PTN PRIMARY KEY (PTN_ID),
 CONSTRAINT       UK_PTN_NUM_CLEF_SECU UNIQUE (PTN_NUM_SECU, PTN_CLEF_SECU),

 CONSTRAINT       CK_PTN_CLEF_SECU CHECK
                  (PTN_CLEF_SECU IS NULL
                     OR (SUBSTRING(PTN_CLEF_SECU FROM 1 FOR 1)
                                 BETWEEN 0 AND 9) AND
                         SUBSTRING(PTN_CLEF_SECU FROM 2 FOR 1)
                                 BETWEEN 0 AND 9)),
 CONSTRAINT       FK_PTN_CVT_ID FOREIGN KEY (CVT_ID) REFERENCES T_CIVILITE_CVT (CVT_ID))
```

Remarquez dans cet exemple que les contraintes ont toutes été exprimées en tant qu'attribut de la table et introduite à l'aide du mot clef CONSTRAINT, et possèdent toutes un nom bien codifié.