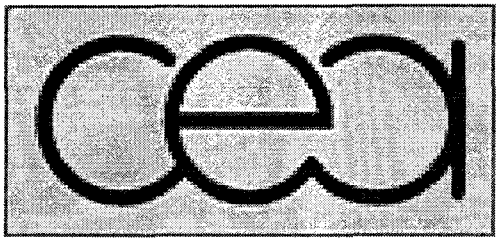
Mise à jour de règles de reformulation

Carole Luczak



Commissariat à l'Energie Atomique

Carole Luczak

Tuteur de stage : Christian Fluhr

Mise à jour de règles de reformulation

27 juin 1997 - 26 septembre 1997

Maîtrise Sciences de l'Information et de la Documentation

Université Charles de Gaulle - Lille III

Je remercie chaleureusement Christian Fluhr pour l'accueil qu'il m'a réservé au sein de son laboratoire et pour les conseils qu'il m'a prodigués.

Je souhaite aussi faire part de ma reconnaissance à Karine Gurtner pour l'aide qu'elle m'a apporté dans la connaissance de SPIRIT.

Un grand merci également à Frank Legrand pour l'aide qu'il m'a apporté sur la reformulation et la gestion des dictionnaires.

Je souhaite également exprimer ma reconnaissance à Jean-Luc Simoni qui m'a longuement parlé de la linguistique pour l'informatique.

Et enfin, merci à toutes les personnes du S.I.I.A. qui m'ont apporté leur aide d'une manière ou d'une autre.

SOMMAIRE

1. Introduction

Mon stage s'est déroulé dans le département de la Section Méthodes et Technologies de l'Information au sein de la Direction de l'Information Scientifique et Technique du Commissariat à l'Energie Atomique (CEA) de Saclay. La DIST étant en cours de restructuration, l'organisation ainsi que l'appellation de ses différents services est amenée à changer prochainement. Le sujet du stage était « mise à jour des règles de reformulation multilingues dans le cadre de la mise en place sur Internet de la base des publications du CEA », Monsieur Fluhr Christian en était le tuteur. Ce stage a duré trois mois de début août à fin septembre.

Le C.E.A. a donc le projet de mettre sa base de publications sur Internet par l'intermédiaire de SPIRIT W3. Internet étant un réseau international il s'est avéré nécessaire de permettre une interrogation multilingue, ce qui implique alors l'existence de règles de reformulation multilingues. L'interrogation de la base des publications du C.E.A. sera bilingue. On devra pouvoir interroger aussi bien en français qu'en anglais.

Mon travail a donc consisté en la traduction anglais-français et français-anglais de mots qu'il fallait ensuite organiser sous la forme stricte des règles de reformulation. Dans ces règles de traduction doivent apparaître aussi bien les différents sens du mot en entrée que les synonymes.

Avant de commencer le travail précisément il a fallu que j'étudie le fonctionnement général de SPIRIT et plus particulièrement celui des règles de reformulation. Pour effectuer le travail, j'ai ensuite utilisé principalement des dictionnaires monolingues et bilingues.

Dans ce rapport je présenterai tout d'abord l'entreprise et le service dans lesquels j'ai effectué mon stage. Je m'intéresserai ensuite au fonctionnement général du système SPIRIT et à celui des règles de reformulation. Je présenterai enfin de façon plus détaillée le travail effectué lors de ce stage.

2. Présentation de l'entreprise et du service

2.1. Le Commissariat à l'Energie Atomique

Le Commissariat à l'Energie Atomique (CEA) créé en 1945 sous l'influence du Général De Gaulle et de Frédéric Joliot est un organisme public de recherche chargé de donner à la France la maîtrise de l'atome dans les secteurs de l'énergie, de l'industrie , de la recherche, de la santé, de l'environnement et de la défense.

Le CEA est une force de proposition, d'expertise et de conseil pour les pouvoirs publics. Il prépare l'avenir et apporte son soutien à l'industrie nucléaire. Il étudie des palettes de solutions scientifiques et techniques afin que les décideurs , pouvoirs publics et industriels, soient en mesure de prendre, en toute connaissance de cause , les solutions les mieux adaptées pour le présent et pour l'avenir.

Les projets nucléaires du CEA sont :

- La contribution à la prolongation de la durée de vie du parc électronucléaire actuel, au développement des réacteurs à eau sous pression de nouvelle génération et à l'accroissement des performances des combustibles, en étroite collaboration avec EdF , Framatome et Cogema (Compagnie générale des matières nucléaires).
- L'étude de nouveaux procédés d'enrichissement de l'uranium avec COGEMA, le retraitement et le recyclage du plutonium, la gestion des déchets radioactifs, en particulier

ceux de haute activité et à vie longue, avec Cogema et Andra (Agence nationale pour la gestion des déchets radioactifs).

• Les recherches et l'expertise en sûreté nucléaire notamment pour le compte de la Direction de la sûreté et des installations nucléaires (DSIN), la protection contre les rayonnements ionisants dont la radiobiologie est l'une des composantes principales et la médecine nucléaire.

• La métrologie des rayonnements ionisants et l'application des radioéléments.

• La fusion thermonucléaire contrôlée.

• Le développement du programme de simulation destiné à maintenir la crédibilité et la fiabilité de la capacité de la dissuasion nucléaire française.

Le CEA a également pour mission, en utilisant les compétences qu'il a développées pour le nucléaire et en coopération avec les autres organismes de recherche, d'apporter sa contribution spécifique aux autres grandes priorités nationales dans différents domaines tels que :

• La recherche fondamentale : physique des particules, physique nucléaire, astrophysique, structures et architectures moléculaires, interaction rayonnement matière, climatologie, biologie cellulaire et ingénierie des protéines.

• Les développements technologiques : micro-électronique, optronique, génie des matériaux, ingénierie des protéines, technologies de l'environnement, instruments pour la recherche scientifique et la diffusion technologique notamment au bénéfice des PME-PMI.

• La transmission du savoir et notamment l'enseignement et la formation par la recherche.

Depuis 1983, l'ensemble des participations industrielles du CEA est détenu par sa filiale à 96%, la holding CEA-Industrie.

Son organisation est simple et décentralisée :

♦ La direction générale

♦ neuf directions fonctionnelles

♦ Sept directions opérationnelles (dont la DIST)

♦ Deux instituts

Le siège du CEA est situé à Paris (XVème). Les activités de recherche et de développement sont réparties sur onze sites.

2.2. La Direction de l'Information Scientifique et Technique

La Direction de l'Information Scientifique et Technique (DIST) est une des sept directions opérationnelles du CEA. Créée le premier mai 1995, la DIST a succédé à la Mission d'Information Scientifique et technique (MIST). Elle a pour missions essentielles de collecter, gérer, élaborer et diffuser l'information scientifique et technique. Elle en assure la pertinence et en garantit le meilleur accès. Elle sensibilise les chercheurs et les ingénieurs du CEA aux méthodes et aux outils modernes dans ce domaine en évolution très rapide.

Implantée sur les sites de Cadarache, Saclay et Valrho, la DIST est, en priorité, au service des unités du CEA. La politique qu'elle se propose de développer a deux objectifs :

1. répondre aux besoins exprimés, c'est-à-dire :
   - ♦ repérer, rendre explicite et analyser ces besoins en information en y associant étroitement les utilisateurs,
   - ♦ maîtriser les outils et les méthodes au service de l'IST (BD, CD-ROMs, serveurs, réseaux, méthodes pour la recherche bibliographique, veille, gestion de documents, gestion des connaissances),
   - ♦ aider les unités à mettre en place des moyens adaptés à leurs besoins.

2. développer la culture d'information scientifique et technique, c'est-à-dire :
   - ♦ sensibiliser les chercheurs et les ingénieurs du CEA à l'importance croissante de l'IST, par exemple en rendant la recherche d'informations plus simple, plus rapide et plus pertinente,
   - ♦ développer les flux d'information élaborée afin de valoriser les compétences et le savoir des équipes du CEA à l'intérieur (mieux exploiter la pluridisciplinarité) et à l'extérieur (offres de service, publications).

Structure de la DIST :

La DIST de Saclay est actuellement en cours de restructuration et aucun des changements n'ayant été rendu officiel nous présenterons l'ancienne structure :

⇒ l'Echelon de Direction

⇒ la Section des Bibliothèques (SBi)

⇒ la Section Collecte d'Information et Bases de Données (SCIBD)

⇒ la Section Recherche et Elaboration de l'Information (SREI)

⇒ la Section Méthodes et Technologies de l'Information (SMTI)

La Section Méthode et Technologies de l'Information :

Elle a pour mission d'étudier et de gérer les techniques de documentation électronique. Le service est ainsi chargé de développer et d'analyser de nouvelles techniques documentaires comme l'hypertexte (liens dynamiques entre différents textes), l'interrogation documentaire en langage naturel.

Dans l'organisation à venir la SMTI et la SREI fusionneraient pour former le Service Ingénierie de l'Information et ses Applications (SIIA).

3. Présentation du logiciel SPIRIT

SPIRIT est un logiciel de gestion documentaire issu des travaux de recherche conduits par A. Andreewsky, F. Debili et C. Fluhr.

Comme tout système documentaire son rôle principal est de satisfaire le besoin d'information exprimé par l'utilisateur.

3.1. Principes généraux

SPIRIT est un logiciel existant en deux versions : une version monoposte ou une version client/serveur. Les serveurs SPIRIT fonctionnent sous deux types d'environnement : l'environnement UNIX ou celui de Windows NT.

3.1.1. Mécanismes d'accès et de recherche de SPIRIT

Nous allons décrire les différents modes d'accès aux documents stockés dans une base SPIRIT.

On distingue les trois modes suivants, disponibles dans SPIRIT

La consultation :

C'est le fait d'accéder à un document dans la base, au moyen de la référence de ce document.

L'interrogation :

C'est la recherche proprement dite. Le résultat de la recherche, effectuée à partir d'une question entrée par l'utilisateur, se compose d'un certain nombre de classes de documents .

Ces classes sont triées par ordre décroissant de pertinence et contiennent chacune un ou plusieurs documents-réponse.

L'accès hypertexte :

Il permet d'accéder directement à un document à partir d'un autre document, à l'aide d'un lien hypertexte appelé renvoi interne.

3.1.2. Recherche par interrogation

Dans ce type de recherche, l'utilisateur saisit une question comportant un ou plusieurs critères de type factuel et/ou textuel. Chacune des zones de saisie d'une grille permet d'interroger simultanément le contenu d'un ou plusieurs champs de la base SPIRIT correspondante, et qui sont toujours de même type. Le type du ou des champs détermine ainsi le type de la zone correspondante.

Pour chacun des critères entrés dans les différentes zones de saisie de la grille courante, SPIRIT examine les portions des index de la base qui correspondent à ces champs. Les opérations d'indexation effectuées sur le contenu des zones textuelles des documents, lors de leur intégration, sont également exécutées sur les critères des questions.

Chacun des critères entrés fait donc l'objet d'un traitement qui dépend de son type, et qui permet d'obtenir :

La liste des documents qui y répondent, pour un critère factuel donné.

La liste des classes de documents qui y répondent (plus ou moins précisément), triée par ordre de pertinence décroissante, pour un critère textuel donné.

Toutes les listes ainsi obtenues sont analysées pour en extraire la liste des documents-réponses.

Traitement des critères factuels

plusieurs opérations sont effectuées successivement par le service d'analyse factuelle de SPIRIT :

Décomposition du critère en conditions élémentaires.

Pour chacune des conditions élémentaires , détermination de la liste des documents qui répondent à cette condition. Cette liste est établie à partir des listes d'index du ou des champs correspondant à la zone de saisie du critère.

Intersection des listes de documents qui répondent à chacune des conditions élémentaires.

traitement des critères textuels

Le traitement de l'interrogation en langage naturel est constitué de plusieurs étapes qui ont une forte similitude avec les opérations d'indexation du contenu d'un champ textuel. Nous allons détailler ces différentes étapes au chapitre suivant.

3.2. Architecture générale

Le système SPIRIT permet de rechercher de l'information à partir de requêtes en langage naturel dans du texte intégral. Il existe deux traitements : le premier est destiné à l'indexation de la base textuelle, le deuxième est destiné à l'interrogation en langage naturel. Les deux traitements utilisent la même analyse morpho-syntaxique soit pour l'extraction des mots significatifs du texte intégral lors de l'indexation, soit pour l'analyse de la requête en langage naturel.

Le schéma ci-dessous nous montre de façon plus détaillée les deux traitements du système SPIRIT qui sont alors appelés « administration de la base » et « consultation de la base ».

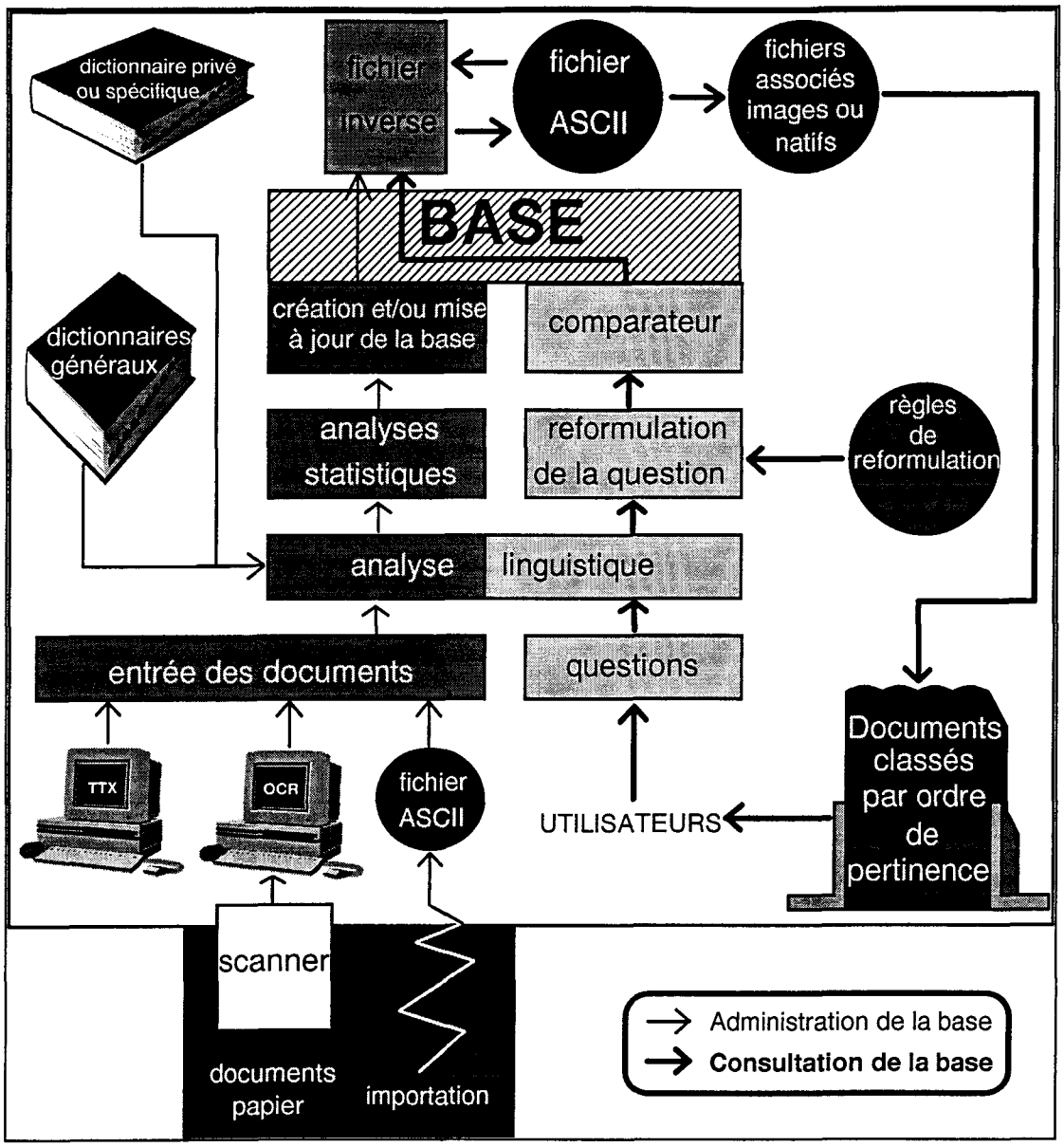


Figure 1 : Schéma de fonctionnement

3.2.1. les différents modules de SPIRIT

Avant de nous intéresser en détail au module principal qu'est l'**analyse linguistique**, passons brièvement en revue les autres modules.

Les **analyses statistiques** se font après l'analyse linguistique lors de l'indexation de la base textuelle. A l'aide d'un modèle statistique elles attribuent un poids pour chaque mot simple ou composé à partir de l'ensemble des mots significatifs. La pondération dans le système SPIRIT est fondée sur le principe selon lequel, plus les mots se trouvent dans de nombreux documents, plus leur poids informationnel est faible.

Contrairement aux systèmes de recherche documentaire booléens, SPIRIT préfère renvoyer les réponses les plus proches plutôt que de ne pas fournir de réponses (notion de silence).
Les poids ainsi calculés par les analyses statistiques sont utilisés lors de l'interrogation, précisément pour le module de comparaison question-documents (ou module **comparateur**), pour calculer la proximité sémantique et trier les documents pertinents selon un ordre décroissant de pertinence.
Le **fichier inverse** sert à mettre en correspondance les différents mots de la base avec les différents textes. Sans ce fichier le système serait obligé de balayer tous les textes à chaque interrogation.
La **reformulation** quant à elle sert à améliorer les résultats de la recherche documentaire en favorisant le rapprochement des concepts exprimés dans la requête à ceux exprimés dans la base. L'action de la reformulation est ainsi destinée à la diminution du silence. Pour parvenir à cet objectif, il faut enrichir les concepts exprimés dans la requête. La méthode adoptée par SPIRIT est orientée vers l'utilisation des règles de reformulation. La partie gauche de la règle contient le mot de la requête, la partie droite contient les mots reformulés. Nous retrouverons des relations comme les synonymes, les termes associés. La reformulation sera l'objet d'un développement plus grand au chapitre suivant.

3.2.2. Les différentes phases de l'analyse linguistique (ou analyse morpho-syntaxique)

L'analyse linguistique est utilisée lors des deux traitements : L'indexation de la base textuelle et l'interrogation en langage naturel. Les différentes phases s'articulent ainsi :

Le découpeur

Le premier problème qui se pose est celui du découpage du texte en mots. Ce travail est réalisé par un automate capable de prendre en compte diverses situations spécifiques. Le point par exemple, qui normalement est un caractère séparateur ne doit pas l'être lorsqu'il fait office de séparateur décimal à l'intérieur d'un nombre. Le résultat de ce découpage est ensuite envoyé à l'analyseur morphologique.

L'analyseur morphologique

Le rôle de cette phase du traitement est de rechercher dans un dictionnaire, pour chaque forme isolée lors du découpage, tout un ensemble d'informations qui s'y rapporte. Il existe pour cela un dictionnaire de formes complètes (full form dictionary). Parmi les informations associées à chacune des formes figurent : la VG (valeur grammaticale), le GN (genre nombre) pour les

adjectifs et les substantifs, le TMP (temps mode personne) et les CVB (caractéristiques verbales) pour les verbes.

C'est à ce stade que se fait l'identification des noms propres, des mots significatifs et des mots grammaticaux ou mots vides. On reconnaît alors aussi les locutions ou expressions idiomatiques. Chaque locution est considérée comme un seul mot insécable dans SPIRIT dans la mesure où le sens de cette locution ne peut pas se déduire simplement de toutes ses parties. Du point de vue documentaire, les locutions sont des mots dont il n'est pas souhaitable que les parties soient un critère de recherche, exemples : « or noir », « chemin de fer », « pomme de terre ». Les locutions sont mises dans le lexique des expressions idiomatiques.

Pour les mots significatifs, les informations relatives à la normalisation sont également extraites à ce stade. Ainsi les adjectifs sont normalisés dans leur forme masculin singulier, les substantif dans leur forme singulier et les verbes dans leur forme infinitive. Le dictionnaire full form qui contient toutes ces informations constitue la base de connaissance du système.

La constitution d'un tel dictionnaire, comportant une entrée pour chaque forme dérivée (450000entrées en français, 150000 en anglais) et contenant autant d'informations grammaticales, représente un travail considérable ; Elle n'est pas dépendante du domaine.

Les informations du dictionnaire full form sont de nature suivante :

Exemples :

Pour le mot « maison » :

VG : S (substantif)

GN : FS (féminin singulier)

NO : maison (normalisé par « maison »)

Pour le mot « magnifiques » :

VG : ADJP ou ADJA ou AT (peut être adjectif postérieur, adjectif antérieur ou attribut)

GN : M ou f, P (peut aussi bien être masculin que féminin, est pluriel)

NO : magnifique (normalisé par « magnifique »)

Il faut noter que lorsque le système traite des textes en typographie non accentuée, c'est lors de cette phase que les mots sont réaccentués. Il convient de remarquer que cette opération n'est pas aussi bénigne qu'il pourrait sembler : l'absence d'accents constituant en fait un manque d'information, la réaccentuation va parfois générer des ambiguïtés. Dans ce cas le

système d'analyse morphologique proposera diverses accentuations possibles . Il sera ensuite du ressort de l'analyseur syntaxique de lever l'ambiguïté dans les cas où cela est possible.
Les homographies sont courantes dans la langue, et la conséquence est un grand nombre de formes ambiguës. Ce sera autant de termes possédant plusieurs séries d'informations, chacune correspondant à une catégorie grammaticale possible. Il faut noter que le faible nombre de dérivation des mots dans une langue comme l'anglais a tendance à favoriser l'abondance de ce type d'ambiguïté. C'est la phase d'analyse syntaxique qui va devoir se charger , sinon de les éliminer, du moins d'en réduire considérablement le nombre.

L'analyse syntaxique

La phase d'analyse syntaxique est plus précisément une phase de désambiguïsation grammaticale. Elle repose sur le principe de compatibilité entre valeurs grammaticales se succédant. L'analyseur utilise dans un premier temps des matrices de compatibilité de successions ternaires de catégories grammaticales. Son travail consiste donc à vérifier que les successions trois à trois des valeurs grammaticales sont acceptables dans la syntaxe de la langue. Un grand nombre d'ambiguïtés peut être levé à ce stade.
L'élaboration des matrices est effectuée par apprentissage lors d'une phase préliminaire à l'exploitation. Celles-ci sont dépendantes des textes d'apprentissage en ce sens que l'on ne peut s'assurer que tous les cas de figure pouvant exister dans la langue y soient présents. La conséquence est qu'il peut se produire des ruptures de séquence au niveau des matrices ternaires. Afin de pallier ce problème, et de pouvoir effectuer la désambiguïsation, le système utilise alors des matrices de compatibilité binaire pour rétablir la continuité.
Ces matrices sont dépendantes de la langue traitée.

Exemple :
Soit la phrase « il ferme le robinet ».
Ici l'ambiguïté sur « ferme » qui peut être substantif ou verbe est levée par la compatibilité de succession car « il » est un pronom sujet, or un substantif ne peut pas succéder à un pronom. C'est donc l'indicatif qui sera retenu comme valeur grammaticale de « ferme » et sa normalisation sera « fermer ».

En sortie, après l'analyse syntaxique, pour la plupart des mots du texte correspond de manière sure une forme normalisée. A cette forme sont associées les informations grammaticales

concernant le mot sous sa forme dans le texte et relativement au contexte syntaxique. Les ambiguïtés résiduelles sont conservées associées aux différentes solutions possibles.

C'est aussi à ce stade que le système procède à l'élimination des mots vides à l'aide des critères syntaxiques et à la reconnaissance des mots composés. Les mots composés ou multitermes sont les termes qui sont en relation de dépendance syntaxique. Le mot composé est significatif pour la recherche documentaire toutefois, la relation entre le mot composé et ses parties étant une relation sémantique pouvant présenter un intérêt pour la recherche documentaire, il y a souvent généralisation. Par exemple, « traduction automatique » est généralisé par « traduction ».

4. La reformulation

Il existe deux sortes de reformulation. La reformulation monolingue qui ne concerne qu'une seule langue à la fois et dont le rôle est d'améliorer les résultats de la recherche documentaire en favorisant le rapprochement des concepts exprimés dans la requête de ceux exprimés dans la base. La deuxième reformulation est celle multilingue dont il existe plusieurs définitions. Nous garderons celle appliquée dans SPIRIT. La reformulation multilingue permet de traiter une collection de documents monolingues dont la question peut être écrite dans différentes langues. Cela permet aux utilisateurs d'exprimer leur besoin d'information dans leur langue maternelle pour interroger une base documentaire écrite dans une autre langue qu'ils peuvent lire.

4.1. Ce qui est à l'origine de la reformulation

4.1.1. En ce qui concerne la reformulation monolingue

C'est la complexité du langage naturel qui est à l'origine de la reformulation. Nous pouvons citer les principaux phénomènes du langage naturel qui posent problème lors du traitement automatique.

La synonymie

Définition : « sont appelés synonymes les unités lexicales de même contenu sémantique mais de forme différente ».

La relation de synonymie est une relation sémantique liant deux mots lexicaux à un même sens. Cette relation de synonymie est dans la plupart des cas non symétrique. Il existe deux sortes de synonymie :

♦ La synonymie absolue : les synonymes absolus sont interchangeables quelque soit la phrase utilisée et sans changement de sens. La relation est alors symétrique. C'est le cas des termes scientifiques, en médecine par exemple :

« lithiase »= « calcul rénal » ;

♦ La synonymie relative : la synonymie relative n'est pas transitive. Elle peut dépendre du contenu sémantique du texte comme elle peut dépendre de la connaissance pragmatique du contexte d'énonciation.

La paraphrase

La paraphrase est un cas particulier de la synonymie. Elle consiste à exprimer une idée de plusieurs manières. La paraphrase est plus difficile à cerner car c'est un phénomène de discours alors que la synonymie est un fait de la langue qu'il est possible de répertorier dans les dictionnaires.

Exemple :

Il cherchait la solution.

Il essayait de trouver la solution.

La polysémie

Définition : « la polysémie est la propriété d'une unité lexicale de posséder plusieurs acceptions ».

Exemple : le verbe *visiter* possède plusieurs acceptions :

1. se rendre auprès de quelqu'un pour le réconforter (*visiter les malades*).
2. Se rendre en un lieu et le parcourir en l'examinant (*visiter un pays*)
3. Examiner minutieusement (*la douane visite les bagages*).

Le sens du verbe « *visiter* » dépend donc du contexte d'énonciation. La paraphrase engendre donc l'ambiguïté dans la compréhension automatique du langage.

Comme nous l'avons vu au chapitre précédent certains types de reformulation sont résolus par le module d'analyse linguistique du système documentaire. Ce module permet en effet de dégager les mots significatifs les plus importants contenus dans le texte sous une forme moins ambiguë. Cette nouvelle forme permet de réaliser une meilleure comparaison entre le contenu de la question et le contenu des documents. Cependant la reformulation effectuée par les différentes phases de l'analyse linguistique ne permet que le contrôle des mots de la langue du point de vue morphologique et syntaxique. Il faut pourtant ajouter un certain nombre de considérations sémantiques qui permettent de réaliser le lien entre les mots et les sens. C'est à ce moment que sont mis en œuvre les différents mécanismes de reformulation de la question de l'utilisateur. Ces mécanismes permettent de représenter les concepts contenus dans le texte et dans la langue en général pour réaliser le rapprochement entre la sémantique de la question et celle contenue dans la collection de documents.

4.1.2. En ce qui concerne la reformulation multilingue

A l'origine de la reformulation multilingue il y a la mondialisation du réseau Internet qui permet la circulation de l'information dans plusieurs langues et l'accroissement du nombre d'utilisateurs pratiquants des langues différentes. Ceci pose le problème d'accès à une collection de documents écrits dans une langue étrangère. Même si les utilisateurs peuvent comprendre le contenu des documents écrits dans une langue étrangère, il leur est plus facile d'exprimer leur besoin d'information dans leur langue maternelle. C'est à ce besoin que correspond la reformulation multilingue dans SPIRIT.

4.2. Structure des règles de reformulation

Nous avons vu que les mots normalisés de la question peuvent subir une reformulation monolingue ou multilingue. La reformulation monolingue permet de réduire le silence du système en rajoutant les termes synonymes ou les termes associés aux mots de la question. La reformulation multilingue permet, quant à elle, la traduction des termes exprimant le concept de la question d'une langue source vers une langue cible. Ceci permet donc de retrouver des documents écrits dans une langue différente de celle de la question.

Pour réaliser ces inférences, le système utilise des dictionnaires de reformulation monolingue et des dictionnaires de reformulation multilingue contenant les règles d'inférence.

Les règles de reformulation sont divisées en deux parties:

- La partie gauche contient le terme-clé à inférer;
- La partie droite contient un ensemble d'inférences représentant tous les sens que peut prendre le terme-clé.

Les différentes parties de la règle de reformulation:

Le mot-clé
Le séparateur de mot-clé: "!" ou une tabulation
La catégorie grammaticale
Le signe "#"
Le type de relation sémantique
Un espace
La liste des sens: les sens des mots sont séparés par des ";" pour une relation sémantique donnée et une catégorie grammaticale donnée. Pour chaque mot suit la catégorie grammaticale, le séparateur de catégorie grammaticale est "$".

Exemple:

chien S#T dog$S;

4.2.1. Les catégories grammaticales

Les catégories grammaticales utilisées pour les règles de reformulation ne sont pas celles qui sont utilisées dans l'analyse morpho-syntaxique du système SPIRIT. Ce sont des macro-catégories qui sont utilisées regroupant les catégories fines utilisées dans l'analyse linguistique.

Liste des catégories utilisées pour les règles de reformulation

Catégories: S (classe des substantifs)
V (classe des verbes)
D (classe des adverbes)

J (classe des adjectifs)

N (classe des noms propres)

4.2.2. Les relations sémantiques

Relation multilingue:

• (T) Traduction: c'est une relation sémantique qui nous permet de traduire un terme d'une langue source vers une langue cible dans tous les sens possibles. Les sens sont séparés par le ";" .

Relations monolingues

• (S) Synonymie: c'est une relation qui permet de reformuler un terme par ses synonymes dans la même langue. Un terme polysémique sera reformulé par les différents sens possibles séparés par ";".

• (A) Associé: c'est une relation qui nous permet de reformuler un terme par les termes de même famille ou par une relation d'association. Cette association est assez large, nous pouvons ainsi associer pêche et mer.

Famille de mots: c'est un ensemble de mots issus de la même racine qui sont en relation sémantique. C'est le cas d'un verbe, d'un substantif ou d'un adjectif qui en dérive, comme dans expérimenter, expérimentation et expérimental.

Les termes inférés se trouvant dans la partie droite de la règle peuvent être des mots simples, des mots composés ou des locutions. Pour reformuler un mot simple, il suffit d'extraire les termes contenus dans la partie droite des règles de reformulation et de les associer au mot de la question. En revanche, la reformulation des mots composés se fait en deux étapes:

♦ Une reformulation globale similaire à l'inférence des mots simples;

♦ Une reformulation par partie qui tient compte des inférences des mots simples que contient le terme composé.

Exemple:

Pour la question "radiation protection", l'analyseur linguistique de SPIRIT présente ceci à la sortie:

radiation (S);

protection (S);

radiation protection (S).

A ces termes sont associées les règles de traduction suivantes:

radiation ⇒ irradiation$S;rayonnement$S;radiation$S;

protection ⇒ protection$S;sauvegarde$S;

radiation protection ⇒ radioprotection$S;

Une seule règle de reformulation est associée à un terme qui peut prendre plusieurs valeurs syntaxiques. Par exemple le terme "absent" peut être soit adjectif, soit substantif.

Exemple de règle de transfert du français vers l'anglais:

absent J#T absent$J,away$J;missing$J;absent-minded$J,inattentive$J;!S#T absentee$S,missing person$S;

C'est le signe "!" qui sert à séparer les différentes valeurs syntaxiques ici adjectif (J) et substantif (S).

5. Mise à jour des règles de reformulation.

Le C.E.A. a le projet de mettre sur Internet sa base de publications par l'intermédiaire de SPIRIT W3. SPIRIT W3 est une passerelle entre un serveur SPIRIT et un serveur HTTP. Elle permet de se connecter à une base, d'effectuer une interrogation, de consulter la liste des documents-réponses et enfin de consulter les documents. Ainsi, à partir du réseau Internet les utilisateurs peuvent interroger la base des publications du CEA par l'intermédiaire du système SPIRIT. Internet étant un réseau international il s'est avéré nécessaire de permettre une interrogation multilingue, ce qui implique alors l'existence de règles de reformulation multilingues. Celles-ci étant encore incomplètes mon travail a consisté en l'ajout d'un certain nombre de règles. L'interrogation de la base des publications du C.E.A. sera bilingue. On devra pouvoir interroger aussi bien en français qu'en anglais.

Ainsi, j'ai eu deux listes. Une liste de mots en français pour lesquels je devais trouver des règles de traduction en anglais et une liste de mots en anglais pour lesquels je devais trouver des règles de traduction en français. Le travail a donc consisté en la traduction de mots qu'il fallait ensuite organiser sous la forme stricte des règles de reformulation afin qu'elles soient acceptées par le système SPIRIT.

5.1. Les domaines concernés par la base des publications du C.E.A.

Le C.E.A. s'intéresse à des domaines assez variés tels que :

L'informatique ;

L'instrumentation ;

Les sciences et techniques nucléaires ;

Les sciences appliquées ;

La physique ;

Les matériaux ;

L'énergie ;

La chimie ;

Ainsi, les mots contenus dans les listes appartiennent généralement à des domaines spécialisés et les mots relèvent donc du vocabulaire technique. Cependant, on trouve également des mots appartenants au vocabulaire général de la langue. L'étude du domaine auquel chaque mot appartient apparaît nécessaire afin d'aller chercher l'information dans les ouvrages appropriés.

5.2. Les outils utilisés

Les outils utilisés sont principalement les dictionnaires.

Pour la traduction anglais-français et français-anglais :

Le Webster : c'est un dictionnaire d'anglais monolingue. Il a été précieux non pas pour la traduction directement mais plutôt pour vérifier si les formes existaient bien en anglais et aussi pour vérifier leur catégorie grammaticale.

Le Robert : c'est un dictionnaire de français monolingue. Il sert à vérifier que les formes existent bien en français et il sert aussi à vérifier les catégories grammaticales.

Le Harrap's shorter : c'est un dictionnaire bilingue (anglais-français ; français-anglais) qui a donc été utile pour la traduction directement. Il n'est valable que pour les mots appartenant au vocabulaire général de la langue.

Le Ernst : c'est un dictionnaire bilingue qui est plus spécialisé, puisqu'il s'intitule « dictionnaire général de la technique industrielle » . Il a été très utile. En effet, suffisamment spécialisé, il permet de traduire beaucoup de mots. Il sait rester aussi assez général pour

aborder de nombreux domaines, ce qui est fort pratique. Il existe en version anglais-français et également en version français-anglais.

Le dictionnaire des sciences et techniques nucléaires du C.E.A. : dictionnaire de français avec une traduction en anglais. Il a été assez peu utilisé car il est tout de même trop spécialisé pour la base en question.

Le lexique Framatome : c'est un lexique bilingue (français-anglais ; anglais-français) qui est très spécialisé. En effet, s'intéressant au nucléaire, il répertorie plus précisément le vocabulaire relatif aux « réacteurs à eau sous pression ». Utile dans le cadre du C.E.A., il n'a pas été beaucoup consulté à cause de sa forte spécialisation.

Les dictionnaires multilingues INIS de l'Agence Internationale de l'Energie Atomique : les deux dictionnaires faisant entre autre la traduction anglais-français et français-anglais ont été intéressants à consulter.

Autres ouvrages plus ou moins consultés : le glossaire français de l'O.C.D.E. ; Le lexique des techniques de l'ingénieur en français ; Le Larousse en ligne ;

6. Les différents problèmes rencontrés

6.1. Problèmes liés à la traduction dans une autre langue: rapprochement avec la terminologie.

Avant de parler de la terminologie, nous allons définir ce que c'est à l'aide d'un certain nombre de définitions.

Définition de la terminologie par l'ISO : « étude scientifique des notions et des termes en usage dans les langues de spécialités ».

Dans son état actuel, la terminologie est fondée sur un modèle tripartite dont les trois points-clés sont l'objet, la notion et le signe.

La plupart des terminologues fondent leur travail sur la référence à une abstraction mentale de l'ordre du concept et dénommée *notion.* Voici les définitions du terme, de la notion et de l'objet telles qu'elles figurent dans la dernière norme ISO 1087 (1990) consacrée au vocabulaire de la terminologie.

« **Notion** : Unité de pensée constituée par abstraction à partir des propriétés communes à un ensemble d'objets.

NOTE- Les notions ne sont pas liées aux langues individuelles. Elles sont cependant influencées par le contexte socioculturel. »

« **Objet** : Elément de la réalité qui peut être perçu ou conçu.

NOTE- Les objets peuvent être matériels (par exemple : moteur) ou immatériels (par exemple : magnétisme). »

« **Terme** : Désignation au moyen d'une unité linguistique d'une notion définie dans une langue de spécialité.

NOTE- Un terme peut être constitué d'un ou plusieurs mots (terme simple ou terme complexe) et même de symboles. »

Le terme a pour principal avantage la non ambiguïté dans un domaine particulier.

Il ne faut pas négliger en terminologie l'importance d'une prise en compte des liens de sens qui unissent les notions entre elles. Ces liens dénommés liens notionnels ou relations notionnelles, sont de nature à éclairer sur la notion :

« Un domaine (ou une sous-section de domaine) n'est accessible mentalement que si le champ notionnel est structuré, c'est-à-dire s'il constitue ce que l'on appelle un système de notions. Dans cet ensemble, chaque notion révèle ses rapports avec les autres notions. » (Felder 1987)

Le travail que j'ai effectué se rapproche de la terminologie. En effet, les mots que j'ai traité appartenaient pour la plupart à des langues de spécialités. Cependant, SPIRIT étant un système documentaire à vocation générale, il a fallu ajouter aussi les sens généraux, quand ils existaient, aux mots à traiter. Par la méthode, le travail effectué diverge de celui d'un terminologue. En effet, la démarche en terminologie est dite onomasiologique c'est-à-dire que l'on part des concepts pour aller vers les termes. Elle implique une étude systématique des concepts. Ma démarche est inverse. Ayant une liste préétablie de mots, je suis partie des termes pour envisager les concepts et trouver ensuite le terme correspondant dans la langue cible. Ceci correspond davantage à la méthode dite sémasiologique qui part du signe pour aller vers la détermination des concepts. Cette démarche est celle employée en lexicologie qui est l'étude des unités lexicales. Ainsi, tout en s'en démarquant, le travail effectué a aussi des points communs avec la terminologie et la lexicologie.

Le problème lié à l'équivalence des concepts reste similaire.

En terminologie multilingue comme pour toute traduction bilingue ou multilingue se pose le problème de l'équivalence des concepts lorsque l'on change de langue. En effet, la linguistique a depuis longtemps montré que toutes les langues n'approchent pas la réalité de la même façon et que de nombreux problèmes se posent lors de l'établissement d'équivalences. Par exemple en français le mot "mouton" désigne à la fois l'animal et la viande de cet animal que l'on peut manger. En anglais, en revanche, il existe deux mots distincts: "sheep" pour l'animal et "mutton" dans le contexte culinaire.

De même, il n'existe pas nécessairement d'équivalence syntaxique. Ainsi, un adjectif dans telle langue n'aura pas forcement d'équivalent sémantique sous une même catégorie grammaticale.

6.2. Les problèmes liés au système documentaire automatisé

6.2.1. La normalisation

Les règles de reformulation interviennent après le traitement linguistique et donc sur des termes déjà normalisés. Il est nécessaire en partie gauche comme en partie droite des règles que les mots soient normalisés lors de la construction des règles puisqu'il n'y a aucun traitement linguistique sur les règles. Par exemple, les adjectifs seront au masculin-singulier comme nous l'avons vu précédemment :

Dutch J#T hollandais$J ;

En ce qui concerne les substantifs, le féminin est conservé bien sûr :

Dutchman S#T hollandais$S,néerlandais$S ;

Dutchwoman S#T hollandaise$S,néerlandaise$S ;

Lors de la construction des règles, il faut donc garder à l'esprit cette normalisation.

Tous les mots grammaticaux ou mots vides sont aussi à exclure des règles. Ceci complique en général la traduction. Par exemple :

combining S#T récolte moissonneuse batteuse$S ; au lieu de « récolte à la moissonneuse batteuse ».

Ce qui pose problème aussi, c'est lorsqu'un mot se traduit en un mot composé. En effet, le système SPIRIT effectuant sa recherche en texte intégral, il faut être sûr de pouvoir retrouver les différents mots d'une traduction côte à côte. Par exemple :

eigenvalue S#T valeur propre$S ;

Pour une recherche efficace, il faut être sûr de trouver les deux mots significatifs « valeur » et « propre » côte à côte tout en sachant que les éventuels mots grammaticaux auront été enlevés précédemment.

Si l'on pense que seul l'ensemble du groupe de mots fait sens et que le sens séparé des différentes parties n'apporte rien de valable, il faut alors rentrer le groupe de mots dans SPIRIT comme une locution. Il existe en effet un dictionnaire particulier qui contient toutes les locutions. Exemple de locution : pomme de terre.

6.2.2. La symétrie

Il existe plusieurs dictionnaires dans SPIRIT, les différents dictionnaires monolingues(français et anglais pour ceux qui nous intéressent) et les dictionnaires de règles de reformulation. Il est nécessaire qu'il existe une symétrie entre ces différents dictionnaires. Si, par exemple, pour le dictionnaire monolingue français, on choisit comme forme normalisée l'abréviation C.E.A. et que l'on fasse donc pointer Commissariat à l'Energie Atomique vers C.E.A., il faudra s'assurer que dans le dictionnaire des règles de reformulation ce sera bien l'abréviation qui sera conservée comme forme normalisée. En effet, quand il existe plusieurs graphies possibles pour un mot et que le sens est identique, il faut faire pointer les différentes graphies vers une forme choisie. Par exemple, le mot français « alcoyle » qui peut aussi s'écrire « alcoyl » et « alkyle ».

6.2.3. La recherche documentaire avant tout

SPIRIT traite la langue de façon automatique. Cependant, le but de SPIRIT n'est pas de prendre pour modèle la langue académique de façon systématique. En effet, le système prend aussi en compte l'utilisateur, ce qui peut créer des divergences avec la langue académique. Prenons l'exemple des noms propres :

Le « Français » est un nom propre, cependant la plupart des gens le considère comme un nom commun et l'écrivent sans majuscule. C'est pourquoi il a été décidé pour SPIRIT de le considérer comme un substantif. Comme dans tout système documentaire, il est important de prendre en compte l'utilisateur potentiel. De toute façon SPIRIT restitue ou enlève systématiquement les majuscules afin de retrouver la forme correcte.

Même si la base des publications fait référence à des domaines spécialisés, les règles que j'ai dû compléter doivent aussi être valables dans un contexte général. Pour le moment SPIRIT ne fait pas la différence entre les différents domaines. Cependant, dans le but d'une utilisation ultérieure, un procédé a tout de même été mis en place.Il est déjà présent dans la structure des règles de reformulation. Ainsi, chaque domaine différent dans la partie droite de la règle est séparé par un « ; ». Ce sera ensuite à l'utilisateur de choisir le domaine qui l'intéresse. En revanche, au sein d'un même domaine, les différents synonymes sont séparés par une « , ».

7. Conclusion

Ainsi, le travail effectué sur les règles de reformulation m'a permis de mieux connaître le fonctionnement général du système SPIRIT. La mise à jour des règles de reformulation tout comme la mise à jour du dictionnaire full form est un travail long à gérer manuellement. C'est pourquoi des essais de traitement automatique sont mis en place. Pour les règles de reformulation de traduction, un système d'appariement semi-automatique est essayé. Les méthodes d'appariement se fondent sur des corpus multilingues en entrée et produisent en sortie des correspondances entre des segments de texte qui sont en relation de traduction dans l'ensemble du corpus. Un segment représente une unité textuelle pouvant relever de différents niveaux allant du chapitre au mot. Les appariements les plus intéressants sont l'appariement fin des mots et des expressions afin de permettre un apprentissage des dictionnaires de transfert.

Le traitement de la langue est tellement lourd qu'il est préférable, dès que cela est possible, d'utiliser un traitement automatique ou semi-automatique.

♦ ANNEXE A : GLOSSAIRE DES TERMES SPIRIT

Administrateur : Personne chargée de définir les paramètres de configuration du serveur SPIRIT, ainsi que les droits d'accès des utilisateurs aux bases SPIRIT.

Application client/serveur : Application fonctionnant selon une architecture client/serveur.

Architecture client/serveur : Architecture matérielle et logicielle, dans laquelle les traitements demandés par un poste client sont exécutés sur un poste serveur, qui retourne au poste client le résultat de ces requêtes.

Base de données textuelles : Se dit de bases de données dont les constituants essentiels sont textuels, c'est-à-dire des textes écrits en langue naturelle (français, anglais, etc.)

Base SPIRIT : Une base de données SPIRIT est une base de données textuelles et factuelles groupées en documents, eux-mêmes organisés en champs (Voir Base de données textuelles, Document, Champ).

Champ : Dans la définition de la structure d'une base SPIRIT, élément permettant la segmentation homogène des données contenues dans tous les documents de cette base. Chaque champ est associé à une entité documentaire, comme par exemple le sujet, le résumé, le nom de l'auteur ou le texte proprement dit du document.
Tout champ possède un nom et un type, qui permettent de l'identifier et de spécifier les traitements effectués sur les données qui y sont stockées.

Champ factuel : Champ d'une base SPIRIT, destiné à contenir des données de taille variable et de nature factuelle (généralement de type numérique, alphanumérique ou date), qui ne sont pas nécessairement décrites par un texte en langue naturelle. Son contenu n'est pas traité par l'analyse linguistique.
Le résultat d'une interrogation effectuée sur des champs de ce type est une liste non ordonnée de documents qui répondent aux critères saisis.

En général, l'ensemble des champs factuels représente la fiche descriptive du document dans la base. Ces champs sont de nature similaire à celle des champs définis dans les bases de type bibliographique classiques, par exemple les champs auteur, date, destinataire, etc.
Par exemple, on peut avoir un champ factuel Auteur qui contienne le nom de l'auteur du document; ce champ permettra de rechercher un document grâce au nom de l'auteur.

Champ Référence : Champ obligatoire, situé en premier dans la structure de documents d'une base SPIRIT, et qui permet d'identifier de manière unique les documents dans la base, auxquels il sert d'identifiant, c'est-à-dire de nom.
Ce champ est généré automatiquement dans la structure de la base par le programme Création de base.

Champ textuel : Champ d'une base SPIRIT, destiné à contenir des données de taille variable et de nature textuelle (du texte en langue naturelle), et dont le contenu sera analysé par le moteur linguistique de SPIRIT.
Le résultat d'une interrogation effectuée sur des champs de ce type est une liste de classes de documents triée par ordre décroissant de pertinence par rapport aux critères saisis.
Les champs textuels sont destinés à recevoir le texte du document SPIRIT, et à faire l'objet d'interrogations en langage naturel. Il est possible de structurer un document SPIRIT en le segmentant en plusieurs champs textuels.
La typographie d'un champ textuel est définie lors de la création d'une base. On distingue la typographie riche, pauvre, ou mixte (voir ces termes).

Document SPIRIT : Fichier texte constitué de plusieurs champs de données ordonnés, de type factuel ou textuel, et stocké dans une base SPIRIT, dans laquelle il peut être lié à d'autres documents d'une même base par un mécanisme d'hypertexte.

Données associées : Données externes associées à un document donné d'une base SPIRIT par l'intermédiaire d'un lien défini dans ce document; d'où leur nom de données associées. SPIRIT n'effectue aucun traitement sur ces données.

Hypertexte : Mécanisme permettant de passer d'un document à l'autre dans une base SPIRIT par l'intermédiaire de renvois.

Langue de la base : Langue naturelle dans laquelle les documents constituant une base SPIRIT sont écrits.

Mode cadré : Mode de calcul des pages informationnelles d'un document réponse, basé sur le découpage en pages logiques du document.
En mode cadré, les pages informationnelles sont des pages logiques sélectionnées et triées en fonction de leur pertinence.

Mode glissant : Mode de calcul des pages informationnelles d'un document réponse, basé sur la répartition des mots informationnels dans le document.
En mode glissant, les pages informationnelles sont des pages dont la première ligne contient un mot informationnel, et dont la longueur est déterminée automatiquement par le système.

Mot informationnel : Mot contenu dans une base SPIRIT, et qui répond à la question posée.
Les mots informationnels obtenus à partir d'une question sont les mots source des mots significatifs de cette question, les mots dérivés de ces mots source, et les mots inférés, obtenus par l'intermédiaire de la reformulation.

Mot significatif : Mot porteur de sens dans la langue naturelle considérée.
Lors de l'intégration d'un document, tous les mots significatifs de ce document seront indexés, afin de permettre d'effectuer des recherches en langage naturel lors de la consultation.

Mot source : Forme normalisée d'un mot quelconque, appelée également lemme, stockée dans le dictionnaire d'une langue et utilisée pour l'indexation en texte intégral des documents d'une base SPIRIT.
Un mot source permet au moteur linguistique de SPIRIT d'accéder à toutes ses formes dérivées. Par exemple, les formes conjuguées des verbes et les accords des substantifs selon le genre et le nombre constituent des formes dérivées de leurs lemmes respectifs.

Objet externe : Objet inséré dans un document SPIRIT, auquel il est relié par un renvoi externe. Il peut s'agir d'un fichier document de traitement de texte, d'une feuille de calcul, d'une image, d'un son, d'une séquence vidéo, etc.

Objet SPIRIT : Terme générique désignant une information déclarative quelconque, marquée par une balise, que contient un document SPIRIT et qui lui est propre, comme par exemple le champ référence, les champs textuels et factuels, les pages logiques, physiques, et image, les renvois hypertexte et les objets insérés.

Page informationnelle : Partie d'un champ textuel d'un document, calculée dynamiquement pour chaque document en fonction de la question posée, qui contient un ou plusieurs mots informationnels.
Il existe deux modes de calcul des pages informationnelles: mode glissant et mode cadré (voir ces termes). Le mode de calcul retenu est défini par la configuration du logiciel d'interrogation.
Les pages informationnelles sont classées par ordre décroissant de pertinence. Elles permettent de se déplacer, en fonction de la question posée, des parties les plus significatives vers les moins significatives du document.

Page logique : Portion d'un champ textuel ou factuel d'un document, délimitée par un découpage, composée d'un nombre variable de lignes contiguës.
Lors de la consultation de ce document, son découpage en pages logiques permet de déterminer les pages informationnelles en mode cadré (voir ces termes).
Les découpages d'un document en pages logiques et physiques sont indépendants l'un de l'autre, quel que soit le mode de constitution et d'intégration à une base SPIRIT de ce document.
Par défaut, si aucun découpage d'un document n'est effectué, SPIRIT Consultation considère que chaque champ correspond à une seule page logique.

Page physique : Portion d'un champ textuel ou factuel d'un document, délimitée par un découpage, composée d'un nombre variable de lignes contiguës.
Lors de la consultation de ce document, son découpage en pages physiques permet de conserver la trace de sa structure originelle en pages de documentation papier, ou bien de définir un découpage de ce type dans un document quelconque.
Les pages physiques permettent également d'associer, le cas échéant, une page image à chaque page physique.

Les découpages d'un document en pages logiques et physiques sont indépendants l'un de l'autre, quel que soit le mode de constitution et d'intégration à une base SPIRIT de ce document.

Protocole : Ensemble de règles spécifiant la manière dont des ordinateurs peuvent dialoguer sur un réseau.

Renvoi externe : Association d'un objet quelconque à un document d'une base SPIRIT. L'objet est stocké dans un fichier spécifique par le serveur SPIRIT, mais il se peut que ce soit un logiciel externe qui se charge de sa manipulation.

Renvoi interne : Lien statique défini spécifiquement dans un document SPIRIT, permettant à un utilisateur d'accéder directement à un autre document de la même base SPIRIT, ou bien à un autre endroit du même document.

Type de champ : Au sens SPIRIT, on distingue trois types de champs: textuel, factuel et référence. Chaque champ est associé à une entité documentaire, comme par exemple le sujet, le résumé, le nom de l'auteur ou le texte proprement dit du document.
Avant l'intégration, chaque champ est présent au plus une fois dans le document. Le champ référence est obligatoire. L'ordre des champs dans le document doit refléter l'ordre de leur définition lors de la création de la base de données.
Lors de la consultation d'une base, la présence de tel ou tel champ dépend de son inclusion dans une zone de la grille d'affichage courante.

Typographie mixte : Représentation des données d'un champ textuel à l'aide d'un jeu de caractères incomplet, intermédiaire entre les typographies pauvre et riche, composé de majuscules, de minuscules, de symboles de ponctuation et de chiffres.

Typographie pauvre : Représentation des données d'un champ textuel à l'aide d'un jeu de caractères restreint, composé seulement de majuscules, de symboles de ponctuation et de chiffres.

Typographie riche : Représentation des données d'un champ textuel à l'aide d'un jeu de caractères complet, composé de majuscules, de minuscules, de minuscules accentuées, de symboles de ponctuation et de chiffres.

♦ ANNEXE B: REFERENCES BIBBLIOGRAPHIQUES

Etude des performances et amélioration des stratégies de reformulation du système documentaire multilingue EMIR. ELKATEB Faiza. Thèse d'état. Université de Paris VII. 1997.

Vers l'accès multilingue en langage naturel aux bases de données textuelles. RADWAN Khaled. Thèse d'état. Université Paris XI. 1994.

Abrégé de terminologie multilingue. VAN CAMPENHOUDT Marc.Site sur Internet, adresse: http://www.refer.fr/termisti/theoweb1.htm

Les langues spécialisées. LERAT Pierre.PUF.1995.