Un modèle d'indexation pour les documents textuels structurés

Francois Paradis

THÈSE

présentée par

François PARADIS

pour obtenir le titre de DOCTEUR
de l'UNIVERSITÉ JOSEPH FOURIER - GRENOBLE 1
(arrêtés ministériels du 5 juillet 1984 et du 30 Mars 1992)
Spécialité : **Informatique**

Un modèle d'indexation pour les documents textuels structurés

---

Date de soutenance : 7 novembre 1996

Composition du jury :

Président :	M. Jacques COURTIN
Rapporteurs :	M. Patrick BOSC
	M. Patrick SAINT-DIZIER
Examinateurs :	M^{me} Catherine BERRUT
	M. Yves CHIARAMELLA
	M. Gregory GREFENSTETTE

Thèse préparée au sein du laboratoire de
Communication Langagière et Interaction Personne-Système – IMAG
Université Joseph Fourier - Grenoble 1

Je tiens à remercier:

M. Jacques Courtin, professeur à l'université Pierre Mendès-France, qui m'a fait l'honneur de présider ce jury.

M. Patrick Saint-Dizier, chargé de recherche au CNRS, pour avoir bien voulu rapporter ce travail, et qui, par ses critiques constructives, a contribué à la forme finale de la thèse.

M. Patrick Bosc, professeur à l'ENSSAT, pour avoir accepté d'être rapporteur, et pour l'intérêt qu'il a manifesté pour ce travail.

M. Gregory Grefenstette, ingénieur de recherche au RXRC à Meylan, pour son aimable participation au jury.

Mme Catherine Berrut, maître de conférence à l'université Joseph Fourier, qui a dirigé ce travail. Ses remarques pertinentes, ses lectures et relectures attentives du manuscrit («enlève-moi ces parenthèses!»), sa grande patience (vertu qu'elle a aussi tenté vainement de me communiquer), et son amitié, ont été autant d'ingrédients essentiels à la bonne réalisation de ce travail.

M. Yves Chiaramella, professeur à l'université Joseph Fourier, pour son accueil au sein de l'équipe et pour le précieux temps qu'il a consacré à mon encadrement. Son expérience et sa vue globale de notre problématique ont été des atouts majeurs pour l'aboutissement de ce travail.

Les membres de l'équipe MRIM, et plus particulièrement Marie-France, pour son entrain et ses encouragements, Philippe, qui avait la réponse à toutes mes questions (ou à défaut, la bouffe), Jean-Pierre, qui a quant à lui trouvé la question à toutes mes réponses, ainsi que les autres membres qui contribuent à faire régner la bonne ambiance au sein de l'équipe: Mourad, Nathalie, Iadh et Franck.

Les autres collègues et invités du bâtiment B: Hassen, sans qui les ballades et la cuisine auraient manqué de piquant, Theo («c'est frais»), Francis, Bernard, et j'en passe...

Mes parents, enfin, qui ont su manifester leur soutien et m'entourer de leur affection malgré les kilomètres.

Résumé

La plupart des modèles d'indexation en recherche d'informations sont spécifiques à une application ou à un domaine particulier, et n'exploitent pas toute la richesse des documents électroniques. Le but de ce travail est de définir un modèle d'indexation pour les documents textuels qui tienne compte de la structure et d'autres informations complémentaires au discours. Le modèle proposé comporte deux composantes: le *langage de représentation*, qui définit de façon conceptuelle les informations du document, y compris les index eux-mêmes, et les *règles de dérivation*, qui, reprenant ce langage, permettent de déduire un type particulier d'index, les *thèmes*. L'indexation dans notre modèle ne se contente pas de produire une représentation statique du document, mais elle est aussi dynamiquement liée au processus de correspondance; ainsi, le choix des thèmes, tels que déterminés par les règles, est fonction du document et de l'utilisateur. Notre approche a été validée en deux temps. D'abord, un questionnaire a été soumis à un groupe d'utilisateurs afin de cerner leur processus de dérivation de thèmes. Cette validation *a priori* a permis de démontrer le bien-fondé de nos règles de dérivation. Puis, dans une validation *a posteriori*, le modèle a été implémenté et testé sur une collection de documents SGML. Cette expérimentation a démontré l'applicabilité et la flexibilité du modèle.

Mots-clés: recherche d'informations, modèle d'indexation, extraction de thèmes, représentation électronique de documents, documents textuels structurés.

Abstract

Most indexing models in information retrieval are dedicated to a particular domain or application, and do not exploit the richness of electronic documents. The goal of this work is to define an indexing model for textual documents that includes structure and other complementary information to the discourse. The proposed model consists of two components: the *representation language*, which defines at a conceptual level the information in the document, including the index themselves, and the *derivation rules*, which are based on this language and enable to deduce a particular kind of index, the *themes*. Indexing in our model does not only produce a static representation of documents, but is also dynamically linked to the correspondence process; in this way, selection of themes, as determined by the rules, is a function of the document and the user. Our approach was validated in two steps. First, a questionnaire was submitted to a group of users in order to understand their process of theme derivation. This *a priori* validation showed the validity of our derivation rules. Then, in an *a posteriori* validation, the model was implemented and tested on a collection of SGML documents. This experimentation showed the applicability and flexibility of the model.

Keywords: information retrieval, indexing model, theme extraction, electronic representation of documents, structured textual documents.

Table des matières

Table des figures

Liste des tableaux

Chapitre 1

Introduction

Notre crime est d'être
homme et de vouloir connaître.

Alphonse de LAMARTINE
(*Premières méditations poétiques*)

LA PROBLÉMATIQUE de la Recherche d'Information peut être vue comme la satisfaction d'un besoin en information d'un utilisateur, qui est exprimé par une *requête*, sur un ensemble de documents appelé *collection* ou *corpus* [vR79] [SM83]. Cette problématique a beaucoup évolué ces dernières années, en grande partie dû à l'augmentation du volume et surtout à une plus grande accessibilité des informations disponibles sous forme électronique. Il suffit de comparer comme témoin de ce changement d'échelle la taille d'une des premières collections-test, la CACM, qui comportait moins de 5000 documents, soit 2 méga-octets de données brutes, avec les expérimentations actuelles de TREC [Jon95], qui sont de l'ordre de 2 giga-octets. De même, avec l'avènement du World Wide Web et d'autres hypermédias, on assiste à une transformation du concept même de document. Celui-ci comporte maintenant des images, des sons, voire même des animations; ainsi que des liens vers les différentes parties du document ou vers d'autres documents. Cette augmentation autant en quantité qu'en richesse a les impacts suivants dans le contexte de la recherche d'informations textuelles:

- l'utilisateur ne cherche plus uniquement un pointeur ou une référence sur l'information comme c'était le cas traditionnellement, mais s'attend souvent à ce qu'on lui retourne directement l'information recherchée;
- les techniques traditionnelles ayant une vision «*plate*» des documents ne sont plus adéquates. Si l'on peut s'en contenter quand les documents sont de taille modeste ou qu'ils ne sont qu'une représentation indirecte des ouvrages complets (comme c'est le cas par exemple pour des résumés), ce n'est plus le cas pour des documents en texte intégral ayant une structure complexe;

- le texte n'étant plus la composante unique du document, il faut revoir sa nature et son rôle par rapport aux autres composantes. Cette nature peut être *typée*, c'est-à-dire que certains éléments sont plus ou moins porteurs d'information selon une étiquette (*markup*) assignée par l'encodeur;
- les systèmes ne traitent plus uniquement des collections statiques et homogènes. L'information évolue avec le temps et est encodée sous des formats divers;

Ces points remettent fondamentalement en cause les principes de base des systèmes de recherche d'informations actuels: la représentation des informations y est souvent trop pauvre, et, même s'ils intègrent certaines des informations citées ci-dessus, ils le font de façon trop *ad hoc* pour faire face à une évolution de la collection ou pour être applicables à d'autres collections.

Le cycle de conception-évaluation est lui aussi remis en question. Les systèmes actuels sont typiquement validés quantitativement par des mesures de *précision/rappel*[1], le but ultime d'une bonne indexation étant de produire des termes d'indexation faisant converger ces mesures vers 1. Or les mesures de précision/rappel n'ont plus de sens: le rappel n'est plus mesurable dès lors que les corpus deviennent trop gros, et la précision est difficile à calculer dans des documents structurés où plusieurs parties peuvent répondre à une même requête. Par conséquent, on ne peut plus actuellement définir une stratégie d'indexation et démontrer son bien fondé *a posteriori* par les résultats du système [RHB92].

Nous proposons un modèle d'indexation dont la richesse répond aux points énumérés ci-dessus, et où la notion d'index est définie de façon à permettre l'évaluation *a priori* du modèle. Notre modèle propose également un lien plus étroit entre recherche d'informations d'une part, et linguistique et théories du discours d'autre part. Dans la suite de ce chapitre, nous fixons d'abord notre vocabulaire, puis examinons de plus près la notion de thème, avant d'exposer les grandes lignes de notre proposition.

1.1 L'indexation en Recherche d'Information

Étant donné les larges volumes de données à traiter, il est généralement admis que la recherche d'informations doit pouvoir s'appuyer sur une représentation synthétique des documents qui en résume les informations susceptibles d'être référencées par un utilisateur. Cette représentation est dénommée *index*, et peut être définie comme suit:

Définition 1 *Un* ***index*** *est une représentation synthétique de l'information relative à un document, qui met en évidence sa sémantique en vue d'une requête.*[2]

1. Le *rappel* est le taux de documents pertinents retrouvés par le système par rapport à l'ensemble des documents pertinents pour une requête, alors que la *précision* est la proportion de documents pertinents parmi tous les documents retrouvés par le système.

2. Cette définition est plus générale que celle trouvée dans [vR79], où les index sont restreints à des mots-clés, mais va dans le sens de Salton [SM83], qui voit les index comme des *représentants* du document.

Les index peuvent prendre différentes formes allant de mots simples à des structures sémantiques plus complexes impliquant plusieurs concepts et relations. Ainsi, la figure 1.1 montre trois possibilités d'index pour une même expression: soit par *mots-clés*, par *groupes conceptuels* ou par *arborescence sémantique*. Les mots-clés sont des termes simples extraits du texte après normalisation [vR79]. Les groupes conceptuels ont été proposés dans [Pal90]. Il s'agit de groupements nominaux reflétant la structure syntaxique et les dépendances linguistiques dans le texte. Un seul groupe conceptuel est dérivé de notre exemple. Il a pour terme directeur «*opacité*»; ses termes dépendants étant indiqués entre parenthèses. Des relations syntaxiques peuvent être matérialisées par des mots liens (prépositions ou autres); dans notre exemple, ce sont «*à_det*» et «*de_det*». Enfin, l'arborescence sémantique présentée ici est du type défini dans le système RIME [Ber88]. Le but de cette représentation est d'expliciter les relations sémantiques entre les divers concepts. L'arbre donné en exemple s'interprète comme suit: l'«*opacité*» porte sur une «*alvéole*» et sur le «*lobe moyen droit*». Ces termes peuvent encore être explicités dans RIME; «*opacité*» devient une «*densité*» ayant pour valeur «*augmentation*», alors que «*lobe moyen droit*» devient un «*lobe*» ayant pour valeurs locatives «*moyen*» et «*droit*».

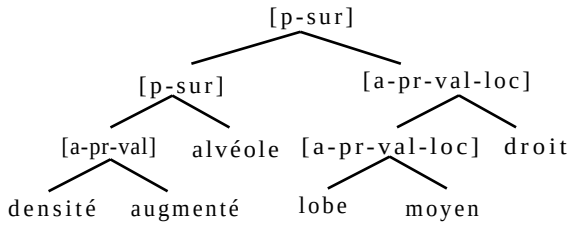
Mots-clés

opacité, alvéol, niveau, lob, moyen, droit.

Groupes conceptuels

opacité (alvéolaire) **à_det** (niveau **de_det** (lobe (moyen) (droit)))

Arborescence sémantique



tiré de [Nie90, p.142]

Figure 1.1. Exemples d'index pour «*opacité alvéolaire au niveau du lobe moyen droit*»

Les index présentés à la figure 1.1 sont exprimés dans des formalismes distincts, chacun possédant ses contraintes syntaxiques, voire même sémantiques: nous dénommons ces formalismes *langages d'indexation*.

Définition 2 *Le* ***langage d'indexation*** *est le formalisme servant à exprimer les index.*[3]

Selon les applications, un formalisme sera plus approprié qu'un autre. Ainsi, une représentation par mots-clés convient pour des collections vastes où les requêtes des utilisateurs portent sur des termes généraux et où la spécificité des liens entre termes n'est pas une nécessité. Par contre, les arborescences sémantiques de RIME ont été définies pour un corpus

3. Cette définition diffère encore une fois de celle proposée par van Rijsbergen dans [vR79], qui considère le langage d'indexation comme décrivant à la fois les documents et les requêtes. Nous préférons dans ce dernier cas parler de langage de requêtes.

médical formé de textes très spécialisés, et où les utilisateurs – des médecins – privilégient la précision par rapport au rappel.

En général, l'*indexation* peut être vue comme un processus de *mise en évidence* de l'information jugée représentative du contenu d'un document.[4] Cette notion d'information représentative est toute relative: elle dépend de l'application, du corpus, des utilisateurs et de leurs besoins en information.

Définition 3 *L'**indexation** (ou le processus d'indexation) est le processus qui permet de construire les index à partir de l'analyse des documents.*

Pour les mots-clés, l'indexation consiste d'abord à sélectionner les termes du document n'apparaissant pas dans un *anti-dictionnaire* et dans une certaine limite de fréquences. Dans notre exemple, les article contractés «*au*» et «*du*» ont ainsi été rejetés. Les termes sélectionnés sont ensuite *lemmatisés*, c'est-à-dire qu'on recherche la racine des mots, le plus souvent en enlevant simplement les suffixes parmi une liste connue. Ceci permet une certaine normalisation des termes dans l'index. Ainsi, dans notre exemple, «*alvéolaire*» est devenu «*alvéol*», ce qui serait aussi le cas pour «*alvéole*». Pour les arborescences sémantiques, l'indexation consiste en une analyse morphologique suivie d'une analyse syntaxico-sémantique.

Un modèle d'indexation, enfin, décrit la représentation des index et le processus de dérivation à partir des documents (définition 4). Il comporte donc deux composantes: le *langage* et le *processus* d'indexation.

Définition 4 *Un **modèle d'indexation** décrit comment représenter des index (*langage d'indexation*) et comment les dériver à partir des documents (*indexation*).*

Beaucoup de travaux se sont intéressés à un modèle général de correspondance[5] [vR86] [Nie90], mais pratiquement aucun ne s'est intéressé à un modèle général pour l'indexation. Les modèles d'indexation dans la littérature correspondent à des règles *ad hoc* qui sont fortement liées à la représentation des documents et à la fonction de correspondance.

Si nous avons tenté autant que possible ici de ne faire aucune supposition quant à l'interprétation des index, il s'avère en pratique que la grande majorité des systèmes de recherche d'informations – même lorsqu'ils ne le disent pas explicitement – considèrent les index comme des thèmes. Malheureusement, si la notion de thème – ou des notions proches – est bien connue en linguistique, elle n'a pas fait l'objet d'études sérieuses du côté de la recherche d'informations. Nous voyons maintenant quelques particularités du thème en recherche d'informations.

4. Il est aussi question de *dérivation* dans [vR79].

5. Le *modèle de correspondance* définit la méthodologie d'évaluation de la correspondance entre requête et documents.

1.2 Le thème en recherche d'informations

En recherche d'informations, le thème fait appel à la notion intuitive de l'«*à-propos*» (*aboutness*) d'un document. Cette notion est plus vague qu'il n'y paraît. Différents mécanismes peuvent intervenir pour sa détermination: la compréhension des idées véhiculées par le document, l'organisation de ces idées en une structure hiérarchique, la connaissance du domaine, etc. La plupart des techniques d'extraction de thèmes tentent d'approximer ces critères par un processus unique; ce faisant, elles risquent toutefois de dépouiller le thème de son sens initial. Par exemple, les techniques traditionnelles, en utilisant des mesures statistiques basées sur la fréquence d'apparition des termes, font la supposition que l'importance d'un thème est uniquement fonction de sa fréquence d'apparition dans le document et la collection.

1.2.1 Thème et contenu sémantique

La figure 1.2 montre le processus de compréhension d'un document du point de vue de la recherche d'informations. L'*auteur* est la personne responsable de la production du document, tandis que le *lecteur* est la personne – ou plutôt le processus puisqu'il s'agit d'un système de recherche d'informations – qui tente de le comprendre. L'auteur a des idées ou des intentions qu'il cherche à transmettre par le biais du document. Le but du lecteur est de retrouver les idées que l'auteur cherchait à exprimer, pour en déduire ce que nous appelons ici les *thèmes*. Pour ce faire, il se bâtit une représentation mentale (le contenu sémantique) du document, et utilise également d'autres informations telles que les connaissances qu'il possède sur ce domaine, la structure du document, etc.

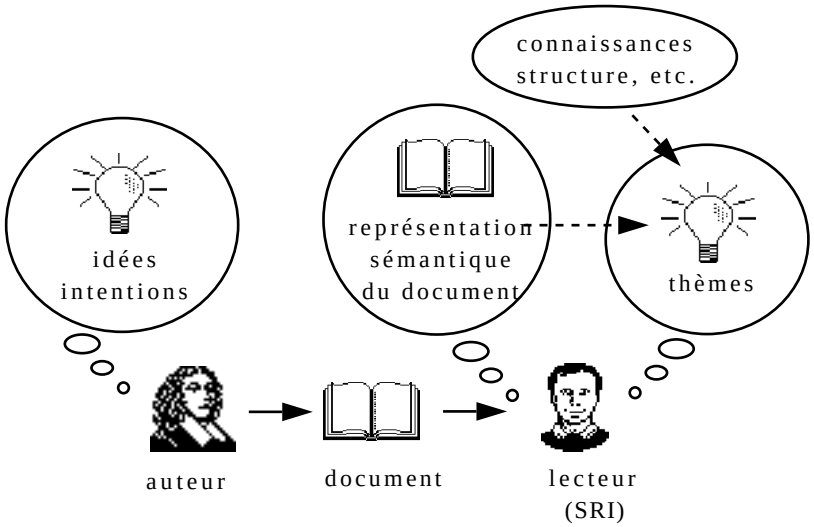


Figure 1.2. Compréhension d'un document

Cette figure illustre aussi la distinction entre le contenu sémantique du document et le thème. Le contenu sémantique est une représentation fidèle du sens du document, où tous les éléments – si peu importants soient-ils – sont représentés, de même que les relations ou contraintes sémantiques entre ces éléments. Le thème, lui, représente les principales idées véhiculées par le document. Par exemple, considérons la phrase «*S'il pleut, Jean apportera son parapluie*». Le contenu sémantique de cette phrase pourrait se traduire par une implication matérielle entre le prédicat *pleuvoir*, qui indique s'il pleut à l'instant t, et le prédicat *apporter*:

$$\forall t \text{ pleuvoir}(t) \supset apporter(Jean, parapluie).$$

Au point de vue du thème, cette relation n'est pas modélisée; cette phrase est à propos de *pluie*, *Jean* et *parapluie*. De même, la proposition «*Il ne pleut pas*» dont le contenu sémantique serait $\neg pleuvoir(t_1)$ (c'est-à-dire, il ne pleut pas à l'instant t_1), aurait aussi pour thème *pluie*.

Ceci ne signifie pas pour autant que toute relation sémantique soit exclue de l'expression du thème. Les thèmes eux-mêmes ont une représentation sémantique; il peut s'agir de simples mots-clés bien sûr, mais aussi de représentations plus poussées incluant des relations sémantiques entre termes simples, comme nous l'avons vu à la section précédente, ou comme il est discuté dans [Par93].

1.2.2 Thème et représentativité

Une autre caractéristique importante du thème est sa *représentativité*.[6] En recherche d'informations, l'intérêt ne porte pas nécessairement sur tout ce qui est dit ou qui est vrai dans un document, mais sur les passages les plus représentatifs ou importants. Ainsi, si le contenu sémantique d'un passage du texte est formé des contenus sémantiques des parties qui le composent, ce n'est pas nécessairement le cas pour le thème. Certains thèmes de sous-sections qui n'apportent rien au thème global, ou dont il n'est pas question dans les autres sous-sections, peuvent être ignorés (bien qu'ils demeurent des thèmes pour leur propre sous-section).

La *représentativité* d'un thème t pour le *passage*[7] d, notée $rep(t, d)$, est fonction de deux critères: sa représentativité *locale* et sa représentativité *globale*, notées *reploc* et *repglob*, respectivement.

$$rep(t, d) = F(reploc(t, d), repglob(t, d))$$

La représentativité locale mesure l'importance de t au sein de d, par rapport aux autres termes qui indexent d. La représentativité globale, quant à elle, est une mesure de l'apport

6. La représentativité du thème est souvent appelée *poids* dans la littérature.

7. Nous entendons ici par *passage* toute sous-partie d'un document ou le document en entier.

de d par rapport à toute la collection. Ainsi, pour la mesure $tf{\cdot}idf$, la représentativité locale est donnée par tf (*term frequency*) ou la fréquence du terme dans le passage, et la représentativité globale, par idf (*inverse document frequency*) ou la proportion du corpus indexée par le terme.

1.3 Principes généraux de notre modèle d'indexation

Nous énonçons maintenant les grandes lignes de notre proposition, quant aux composants du modèle d'indexation, et quant à la définition du thème et de sa représentativité.

1.3.1 Indexation

Nous avons vu que chaque approche définit son propre modèle d'indexation, généralement dépendant du format des fichiers et du domaine d'application, rendant difficile la transposition de l'approche à d'autres contextes. De plus, comme le langage d'indexation se limite à la description des index, d'autres informations pouvant être utilisées pour dériver les index se trouvent «*cachées*» dans le processus d'indexation;

Nous proposons de définir un *méta-langage*, que nous appelons le *langage de représentation* ou $\mathcal{L}$, qui, sans remplacer le langage choisi lors de l'implémentation du modèle, précise à un niveau conceptuel la nature et le rôle des éléments d'information des documents. Ce langage a la particularité d'inclure à un même niveau le *langage d'indexation*, pour la représentation des index, et un *langage de description*, pour la représentation de toute autre information sur le contenu ou à propos des documents.

Le modèle d'indexation est formé de ce langage de représentation et du *processus d'indexation*, exprimé par des *règles de dérivation*, qui permettent d'inférer les index à partir des informations relatives aux documents. Ces deux composants du modèle d'indexation, et leurs liens avec les informations exprimables dans le langage de représentation, sont schématisés à la figure 1.3.

Afin de définir des règles d'indexation concrètes, il est essentiel de se baser sur une interprétation donnée des index. Dans notre travail, comme dans la majorité des systèmes de recherche d'informations, nous avons choisi de nous limiter à la dérivation de thèmes.

1.3.2 Thème

La définition «*philosophique*» du thème est souvent loin de celle impliquée par les techniques qui sont utilisées pour le dériver dans les systèmes de recherche d'informations. Ceci entraîne un décalage entre ce que l'utilisateur croit exprimer ou ce qu'il attend comme réponse, et l'interprétation du système.

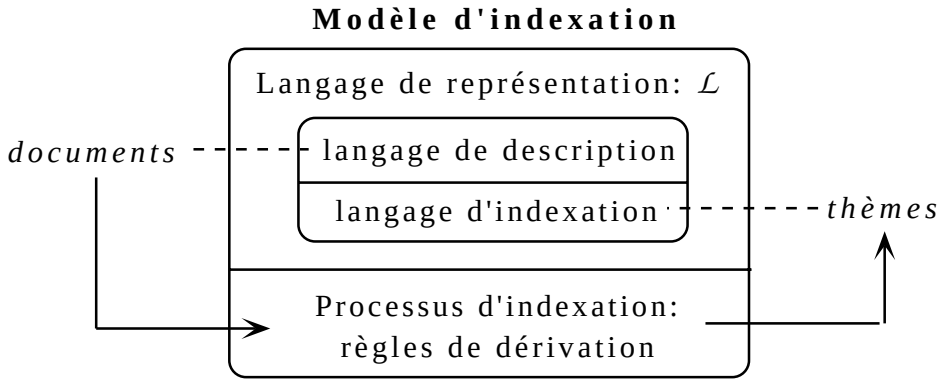


Figure 1.3. Composants de notre modèle d'indexation

Nous proposons de définir les règles de dérivation directement à partir de théories linguistique ou de discours qui reflètent l'interprétation intuitive qu'a l'utilisateur du thème.

Nous avons vu que les mesures de représentativité du thème, étant uniquement quantitatives, ne conviennent pas pour des documents structurés. De plus, étant figées une fois pour toutes lors de l'indexation, elles ne correspondent pas au jugement de pertinence de l'utilisateur, qui est flexible et qui dépend de plusieurs facteurs.

Nous définissons une mesure de représentativité qualitative, qui permet différents jugements de pertinence selon le type de requête. La représentativité d'un terme t pour un passage d est donnée par le couple des représentativités locales et globales:

$$\mathrm{rep}(t,d) = \langle \mathrm{reploc}(t,d), \mathrm{repglob}(t,d) \rangle$$

La représentativité locale correspond à la *justification* d'un thème, ou la raison pour laquelle il a été choisi, et la représentativité globale, à son *contenu sémantique*, c'est-à-dire à l'importance de ce thème pour la collection.

1.4 Organisation de la thèse

La suite de cette thèse est organisée comme suit:

- le chapitre 2 discute de la notion d'*information* dans un modèle d'indexation. L'information est vue ici comme tout élément du document – qu'il soit explicite ou implicite – qui peut être utilisé lors du processus d'indexation;
- le chapitre 3 définit le langage de représentation $\mathcal{L}$, qui décrit les index et les informations du document qui peuvent servir à les dériver;

- au chapitre 4, nous décrivons le processus d'indexation, ainsi que l'expérience qui a permis de valider les règles de dérivation *a priori* auprès d'utilisateurs;
- enfin, au chapitre 5, nous discutons d'une application de notre modèle à un corpus technique, les «*Recommandations TEI*», qui sont sous format SGML.

Chapitre 2

Qu'est-ce que l'information?

Qui n'attend pas l'inattendu ne le découvrira point;
il demeurera hors de portée, indécelable.

HÉRACLITE
(*traduction libre*)

LA NOTION D'INFORMATION , dans son usage courant, est chargée d'ambiguïté. *Le Petit Robert* (édition 85) en discerne six sens différents, alors que le *Webster* (9e édition) en énumère pas moins de neuf. En informatique et dans les sciences cognitives, l'*information* est souvent assimilée aux *données*, et opposée aux *connaissances*, ces dernières étant plutôt vues comme une «*modélisation de l'information dans le but de son utilisation à bon escient*» [Kay84].

En recherche d'informations, malheureusement, le sens d'*information* n'est guère mieux cerné, puisque cette notion n'est jamais formellement définie, et cela aussi bien pour les modèles que pour les systèmes.[1] On se contente de renvoyer à un «besoin de l'utilisateur», qui, bien qu'il exprime indirectement une information, ou plutôt un *manque* d'information, ne peut à lui seul en expliciter la nature. Ce besoin en information est bien souvent thématique. Par rapport à la notion de thème, telle que nous l'avons vue au chapitre précédent, l'information est plus générale, en ce sens qu'elle désigne, en plus du thème, toute information contenue implicitement ou explicitement dans le document qui permette de dériver les thèmes.

Le but de ce chapitre est de cerner la notion d'information dans le cadre de la recherche d'informations textuelle[2], notion qui formera la base de notre modèle d'indexation. S'il est vrai que cette question a fait l'objet de nombreuses études par des disciplines très diverses,

1. Il reste encore à voir si le sens qu'on prête à *information* dans «*recherche d'informations*» est bien le même que dans sa contrepartie anglaise, «*information retrieval*», preuve que nous laissons à d'autres...

2. Nous entendons par recherche d'informations textuelle la classe de systèmes classiques où requêtes et documents portent sur du texte, par opposition aux systèmes multimédia où l'on traite également des images, vidéo, etc.

telles que les sciences de la communication, la linguistique, et les sciences cognitives, pour n'en nommer que quelques-unes, le lien entre ces disciplines et la recherche d'informations demeure largement inexploré. Notre démarche consiste donc à présenter ces divers travaux, pour en dégager ensuite une notion d'information qui réponde aux besoins spécifiques de la recherche d'informations, tout en empruntant aux travaux d'autres disciplines.

Après une brève mise en situation, les travaux appartenant à ce qu'il convient d'appeler la «théorie de l'information» sont d'abord présentés; ils s'agit de travaux de fond s'attaquant à la nature même de l'information. Nous présentons ensuite les travaux de linguistique et de théorie du discours, qui nous permettent de cerner plus précisement à quoi correspond l'information véhiculée par le langage. La question est enfin examinée du point de vue de la recherche d'informations. Nous concluons avec notre proposition du concept d'information pour un modèle d'indexation. Cette proposition tente de concilier les différentes vues de l'information qui ont été exhibées par les travaux précédents, en définissant plusieurs types d'information. Le problème de *mesure* de cette information est aussi traité.

2.1 Classification de l'information

Notre hypothèse de base est que toute information peut être représentée par des symboles, hypothèse qui est d'ailleurs implicite dans la plupart des travaux s'étant attaqué au problème. Trois aspects, qui sont empruntés à la linguistique, peuvent alors être considérés par rapport à ces symboles:

- le *signifiant*, c'est-à-dire la forme matérielle des symboles;
- le *signifié*, qui réfère à la signification qui est véhiculée par les symboles;
- la *pragmatique*, qui s'intéresse aussi à la signification liée aux symboles, mais prise dans un contexte, c'est-à-dire en la situant dans un discours et des connaissances.

Soit par exemple le concept de *parapluie*. Le signifiant s'intéresse au nom commun «*parapluie*», ses origines, son emploi dans une phrase, etc. Le signifié, lui, étudie le concept dénoté par ce mot: il peut s'agir de sa classification par rapport à d'autres types de concepts, de la description physique d'un parapluie, de sa fonction, etc. Enfin, la pragmatique traite de considérations comme la différence dans l'emploi d'un parapluie et d'une ombrelle, le fait que ce sont des objets qui s'oublient facilement dans un autobus, etc.

La table 2.1 présente les travaux dont nous discutons dans les sections à venir, selon ces trois axes. Ces travaux sont aussi groupés selon les trois disciplines déjà mentionnées. Il va sans dire que cette classification est approximative, puisque les frontières sont souvent floues, particulièrement pour ce qui est de la recherche d'informations. De plus, la synthèse des travaux ne saurait être exhaustive: les travaux présentés ont été choisis pour

leur représentativité par rapport à notre travail, et seuls les aspects qui y sont pertinents sont développés. Ces choix apparaîtront plus clairement dans les grandes lignes de notre proposition, à la section 2.5.

Table 2.1. Classification de l'information dans diverses disciplines

	signifiant	*signifié*	*pragmatique*
théorie de l'information	transmission de l'information	information sémantique	théorie des situations
information et langage	morphologie	analyse statutaire progression thématique	intentions théories du discours
recherche d'informations	booléen groupes & surface	concepts	concepts & contexte

2.2 Théorie de l'information

Nous regroupons dans cette section des travaux qui, bien que d'origines diverses, ont contribué à leur manière à une définition générale du concept d'information. Ces travaux sont classés selon les trois aspects déjà cités: les travaux de Shannon-Weaver sur la transmission de l'information (section 2.2.1) correspondent au *signifiant*, ceux de Bar-Hillel sur l'information sémantique (section 2.2.2) appartiennent au *signifié*, et enfin, les travaux de Barwise sur la théorie des situations (section 2.2.3) sont au niveau de la *pragmatique*.

2.2.1 Transmission de l'information

Les travaux de Claude E. Shannon et de Warren Weaver sur l'échange d'information dans un processus de communication [WS75], furent parmi les premiers essais modernes tentant de cristaliser la notion d'information, et font encore aujourd'hui lieu de référence dans les *sciences de la communication*.

La figure 2.1 illustre le principe général d'un processus de communication. Le processus de communication est pris ici dans son sens large; il inclut tous les procédés par lesquels un agent, qu'il soit personne ou machine, peut en influencer un autre. Il ne se limite donc pas à la langue parlée ou écrite, mais comprend aussi d'autres activités comme la musique, la danse, etc.

Une *source d'information* choisit d'abord un *message* parmi un ensemble de messages possibles. Ce message est ensuite encodé par l'*émetteur* en un *signal* qui sera transmis

au *récepteur* par le biais d'un *canal de communication*. Le récepteur à son tour change le signal reçu en message et l'achemine à sa *destination*.

Par exemple, pour une conversation téléphonique, la source et la destination sont les deux personnes qui conversent et le message les paroles qu'elles s'échangent. L'émetteur et le récepteur sont un ensemble d'éléments (émetteur/récepteur téléphonique, etc.) qui transforment la pression du son vocal en un courant électrique variable, et réciproquement le courant électrique en son. Le canal est un fil, et le signal, enfin, est le courant électrique qui parcourt ce fil.

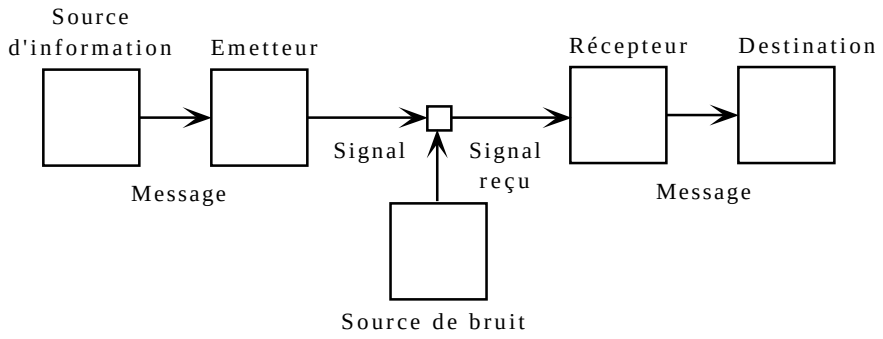


Figure 2.1. Un système de communication (tiré de [WS75, p.35])

Dans cette optique, l'information peut être vue comme «ce qui est transporté ou transmis de l'émetteur jusqu'au récepteur». Au sens de Shannon toutefois, l'information est plus une mesure qu'une entité; elle correspond à la liberté de choix dont on dispose quand on sélectionne un message à transmettre. Cette mesure repose sur des critères statistiques, non pas basés sur une communication particulière, mais sur les probabilités d'utilisation d'un message sachant tous les messages possibles. Elle est formalisée comme suit:

$$H = -\sum p_i \log_2 p_i$$

où H représente la *quantité d'information*, et p_i la probabilité de sélection d'un message i parmi tous les messages possibles.

Dans le cas de (n) messages équiprobables, cette équation devient:

$$H = -\log_2 \frac{1}{n} = \log_2 n$$

ce qui correspond, comme tout bon informaticien l'aura remarqué, au nombre de bits nécessaires pour encoder n messages distincts. Par exemple, étant donné 32 messages équiprobables, on obtient $H = 5$.

Pour un nombre de messages donné, la valeur de H est toujours à son maximum pour des messages équiprobables, et décroit à mesure que certains messages deviennent plus probables que d'autres. Ceci correspond bien à l'idée intuitive de «liberté de choix»; elle est à son maximum quand tous les messages peuvent être choisis indifférement, et à son minimum – à la limite nulle – si un des messages est sélectionné avec une probabilité de 1.

Une constatation intéressante est que l'équation de quantité d'information de Shannon est similaire à la loi de l'entropie de Boltzmann, bien connue en thermodynamique. Le fait de recevoir un message fixe la position relative d'un certain nombre d'éléments par rapport au possible désordre qui aurait pu résulter du tirage de ces éléments au hasard, et par conséquent, le message fournit une certaine information. De même, plus une situation est organisée, plus les choix sont limités, et donc plus l'entropie, donnée par la quantité d'information H, est faible. Ajoutons que l'entropie est aussi assimilée à la notion d'*incertitude* [Ré84], cette dernière étant alors vue comme un *manque* d'information.

Une mesure intéressante qui découle de l'entropie est la *redondance*, donnée par le complément à un de l'*entropie relative* (le rapport de H sur l'entropie maximale pour une même source). Cette mesure quantifie la portion des messages qui est «superflue», c'est-à-dire qui était déjà déterminée par les probabilités d'emploi des messages. Par exemple, le mot «*information*» dans «*recherche d'informations*» peut être jugé superflu, du moins dans notre contexte, puisque «*recherche d'*» sera toujours suivi par «*information*». Les langues naturelles comportent une grande partie de redondance, caractéristique qui est souvent utilisée par les systèmes de reconnaissance de parole comme ceux basés sur les chaînes de Markov. Par exemple, la redondance de l'anglais est estimée à plus de 50% [WS75, p.44].

Si nous revenons maintenant à la figure 2.1, on constate la présence d'une «source de bruit» qui transforme le signal lors de sa transmission. L'interprétation de ce bruit par rapport à l'information qu'il modifie est intéressante. Puisque l'introduction du bruit amène des distorsions, erreurs ou informations ajoutées, on peut en déduire que l'*incertitude* associée au message est accrue, et que donc la quantité d'information l'est également. On peut donc dire que le message reçu dans un système avec bruit contient plus d'information, même si une partie de cette information est falsifiée et indésirable.

Les travaux de Shannon, remarquables par leur généralité, ont trouvé bon nombre d'applications, en informatique notamment avec les théories du codage. Ils ont cependant souvent été qualifiés à tort de «Théorie de l'Information», dont ils ne couvrent en fait qu'un aspect, celui de la transmission de messages entre deux agents. Il serait donc plus approprié de parler de «Théorie de la transmission de l'Information».

Shannon se garde bien de confondre *information* – qui se situe pour lui au niveau du signifiant – et *signifié*. Ainsi, un message fournit d'autant plus d'information qu'il est moins probable, quelle que soit sa signification. Par exemple, le message «*les idées vertes dorment furieusement*», même s'il n'a aucun sens, est porteur de beaucoup d'information de par sa faible probabilité d'apparition.

S'il peut sembler surprenant à première vue de ne considérer que des aspects statistiques

de l'information, il s'avère que ces derniers sont intimement liés à la signification (qu'on pense à la redondance dans les langues naturelles par exemple).

2.2.2 Information sémantique

La théorie de la transmission de l'information présentée à la section précédente néglige délibérément tous les aspects sémantiques, c'est-à-dire relevant de la signification des messages. La théorie de l'information sémantique [BH64], quant à elle, traite de l'information véhiculée par une phrase et de différentes mesures qui peuvent y être appliquées. Contrairement à l'«information-signal» de Shannon, l'information sémantique existe en elle-même, indépendamment du fait qu'elle soit transmise ou non. Les concepts d'*émetteur* et *récepteur* sont donc absents; l'information sémantique ne traite pas de l'information que l'émetteur avait l'intention de transmettre, ou de l'information telle qu'elle a été perçue par le récepteur.

Le concept d'information est associé à celui de *contenu*, noté par *Cont*. Certaines contraintes sur *Cont*, ayant trait à l'inclusion, l'équivalence, etc., sont énoncées dans [BH64], et conduisent à la définition suivante: le contenu d'une proposition i, noté $Cont(i)$, est donné par l'ensemble des propositions qui sont logiquement déduites de i.

Afin d'expliciter cette notion de contenu, supposons un langage simple $\mathcal{L}$ pour lequel les connecteurs habituels – négation, disjonction, conjonction, implication logique et équivalence – sont définis. Définissons également sur $\mathcal{L}$ deux prédicats: M, qui indique si un individu est de sexe masculin, et E, qui indique si un individu est étudiant. Supposons enfin deux individus, a et b.

Un *état-descriptif* dans $\mathcal{L}$ est constitué de la conjonction des prédicats M et E pour chacun des deux individus, indiquant donc leur sexe et leur statut étudiant. Le langage $\mathcal{L}$ possède 16 états-descriptifs distincts, qui sont donnés à la table 2.2. Par exemple, l'état-descriptif n° 7 signifie que a est masculin mais pas étudiant, et que b n'est pas masculin mais est étudiant.

Un *élément-de-contenu* est donné par la négation d'un état-descriptif. Les éléments-de-contenu pour le langage $\mathcal{L}$ sont aussi donnés dans la table 2.2. Par exemple, l'élément-de-contenu associé à l'état-descriptif n° 7 englobe les situations ou l'une (ou plusieurs) des conditions suivantes est vérifiée: a n'est pas masculin, a est étudiant, b est masculin, ou b n'est pas étudiant.

$Cont(i)$ est alors formalisé par l'ensemble des éléments-de-contenu qui sont impliqués logiquement par i:

$$Cont(i) \equiv \{e \text{ est élément-de-contenu } | i \supset e\}$$

Par exemple, toujours dans le langage $\mathcal{L}$, la proposition $M(a) \wedge E(a)$ a pour contenu les éléments-de-contenu n^{os} 5 à 16, c'est-à-dire tous ceux associés aux états-descriptifs où

Table 2.2. États-descriptifs et éléments-de-contenu dans $\mathcal{L}$

n°	état-descriptif	élément-de-contenu
1.	$M(a) \wedge E(a) \wedge M(b) \wedge E(b)$	$\neg M(a) \vee \neg E(a) \vee \neg M(b) \vee \neg E(b)$
2.	$M(a) \wedge E(a) \wedge M(b) \wedge \neg E(b)$	$\neg M(a) \vee \neg E(a) \vee \neg M(b) \vee E(b)$
3.	$M(a) \wedge E(a) \wedge \neg M(b) \wedge E(b)$	$\neg M(a) \vee \neg E(a) \vee M(b) \vee \neg E(b)$
4.	$M(a) \wedge E(a) \wedge \neg M(b) \wedge \neg E(b)$	$\neg M(a) \vee \neg E(a) \vee M(b) \vee E(b)$
5.	$M(a) \wedge \neg E(a) \wedge M(b) \wedge E(b)$	$\neg M(a) \vee E(a) \vee \neg M(b) \vee \neg E(b)$
6.	$M(a) \wedge \neg E(a) \wedge M(b) \wedge \neg E(b)$	$\neg M(a) \vee E(a) \vee \neg M(b) \vee E(b)$
7.	$M(a) \wedge \neg E(a) \wedge \neg M(b) \wedge E(b)$	$\neg M(a) \vee E(a) \vee M(b) \vee \neg E(b)$
8.	$M(a) \wedge \neg E(a) \wedge \neg M(b) \wedge \neg E(b)$	$\neg M(a) \vee E(a) \vee M(b) \vee E(b)$
9.	$\neg M(a) \wedge E(a) \wedge M(b) \wedge E(b)$	$M(a) \vee \neg E(a) \vee \neg M(b) \vee \neg E(b)$
10.	$\neg M(a) \wedge E(a) \wedge M(b) \wedge \neg E(b)$	$M(a) \vee \neg E(a) \vee \neg M(b) \vee E(b)$
11.	$\neg M(a) \wedge E(a) \wedge \neg M(b) \wedge E(b)$	$M(a) \vee \neg E(a) \vee M(b) \vee \neg E(b)$
12.	$\neg M(a) \wedge E(a) \wedge \neg M(b) \wedge \neg E(b)$	$M(a) \vee \neg E(a) \vee M(b) \vee E(b)$
13.	$\neg M(a) \wedge \neg E(a) \wedge M(b) \wedge E(b)$	$M(a) \vee E(a) \vee \neg M(b) \vee \neg E(b)$
14.	$\neg M(a) \wedge \neg E(a) \wedge M(b) \wedge \neg E(b)$	$M(a) \vee E(a) \vee \neg M(b) \vee E(b)$
15.	$\neg M(a) \wedge \neg E(a) \wedge \neg M(b) \wedge E(b)$	$M(a) \vee E(a) \vee M(b) \vee \neg E(b)$
16.	$\neg M(a) \wedge \neg E(a) \wedge \neg M(b) \wedge \neg E(b)$	$M(a) \vee E(a) \vee M(b) \vee E(b)$

$\neg M(a)$ ou $\neg E(a)$ (ou les deux) apparaissent.

Cette définition de contenu peut aussi être vue de façon plus intuitive comme l'ensemble des états-descriptifs qui sont exclus par une proposition. Ainsi, une tautologie comme $M(a) \vee \neg M(a)$ a un contenu nul, puisque qu'elle n'exclut aucun état-descriptif.[3] Par contre, une proposition se contredisant, telle que $M(a) \wedge \neg M(a)$, a quant à elle un contenu maximum, puisqu'elle exclut tous les états-descriptifs.

Il est facile de comparer les informations entre elles dans le cas d'inclusion simple. Par exemple, soit $i = M(a)$, et $j = M(a) \vee E(b)$. Puisque l'ensemble des éléments-de-contenu de i (les n^{os} 9 à 16) inclut tous les éléments-de-contenu de j (les n^{os} 10, 12, 14 et 16), il est clair que $Cont(i) \supset Cont(j)$.

En général cependant, afin de pouvoir traiter des informations qui n'ont aucun lien entre elles, une mesure du contenu doit être définie. Les propriétés que devrait posséder une telle mesure sont définies dans [BH64]. Certaines découlent directement de la définition

3. Si ce résultat peut sembler surprenant à prime abord, il est tout à fait en accord avec la définition du dictionnaire de «tautologique»: « *redondant, qui n'apporte aucune information* ».

de *contenu*, alors que d'autres sont plus subjectives, c'est-à-dire qu'elles pourront être désirables ou non-désirables selon le cas. Bar-Hillel en déduit qu'une définition unique de mesure du contenu est impossible, et propose deux mesures, *cont* (avec un «*c*» minuscule) et *inf*, ayant chacune leur interprétation.

La première mesure de contenu est définie comme suit:

$$cont(i) \equiv 1 - m(i)$$

où $m(i)$ est la probabilité logique de l'énoncé i. Plus cette probabilité est élevée, et plus petite est la mesure du contenu.

Voyons maintenant quelques-unes des propriétés qui sont vérifiées pour *cont*. La propriété 2.1 stipule que la mesure de contenu d'une information est nulle si le contenu de cette information est aussi nul. C'est le cas des tautologies, comme nous l'avons déjà mentionné. La propriété 2.2 reprend quant à elle l'inclusion simple entre informations.

$$cont(i) = 0\,;\ \text{si}\ Cont(i) = \emptyset \tag{2.1}$$

$$cont(i) > cont(j)\,;\ \text{si}\ Cont(i) \supset Cont(j) \tag{2.2}$$

La notation $i|j$ réfère à l'information i relativement à l'information j (ou l'information i *sachant* j). Ainsi, la propriété 2.3 signifie que l'augmentation de l'«évidence» n'augmente pas la quantité d'information.

$$cont(i|j) \leq cont(i) \tag{2.3}$$

La notation $i\&j$ réfère à l'union de i et j. La propriété 2.4 énonce la compositionalité de l'information. Soit deux informations de contenu disjoint: alors la mesure de leur contenu par union est égale à la somme de leurs mesures de contenu prises séparément. Cette propriété est aussi appelée l'*additivité*.

$$cont(i\&j) = cont(i) + cont(j)\,;\ \text{si}\ i\ \text{et}\ j\ \text{ont un contenu disjoint} \tag{2.4}$$

La seconde mesure, *inf*, s'énonce comme suit:

$$inf(i) \equiv -\log_2 m(i)$$

Cette mesure vérifie certaines des propriétés de *cont*, mais pas toujours sous les mêmes conditions. Les propriétés 2.1 et 2.2 sont toujours vérifiées, par contre la propriété 2.3 n'est plus vérifiée. La propriété 2.4 (*additivité*), n'est vérifiée que pour les propositions

logiquement indépendantes, c'est-à-dire ces propositions dont les probabilités respectives n'ont pas d'influence l'une sur l'autre.

$$inf(i\&j) = inf(i) + inf(j) \ ; \ si \ i \ et \ j \ sont \ logiquement \ indépendantes \tag{2.5}$$

Un exemple amusant inspiré d'un roman policier, et qui illustre bien la difficulté d'une définition de mesure de contenu sémantique, est donné dans [BH64]. Supposons les trois suspects A, B et C d'un meurtre. Deux témoins, X et Y, peuvent innocenter chacun un des suspects. Le procureur chargé de l'affaire promet une récompense pour toute information amenant à l'identification du meurtrier. Il reçoit le premier jour X, qui prouve que A n'a pas pu commettre le crime. Le second jour, Y fait de même pour B. Le procureur en déduit donc que C est le coupable. La question est maintenant de savoir comment partager la récompense entre X et Y. Doit-elle être partagée également? Doit-on donner plus à Y, puisque son information éliminait une personne sur deux, alors que X n'éliminait qu'une personne sur trois? Ou doit-on tout donner à Y, puisque c'est lui qui a permis l'identification finale du meurtrier?

Les liens entre la théorie de l'information sémantique et la théorie de la transmission de l'information sont multiples. Tout comme Shannon ne s'intéressait pas au *sens* de l'information, dans cette théorie, l'information sémantique n'est pas associée à une valeur de vérité. Une proposition fausse, mais qui «dit» beaucoup, est considérée comme très informative. Si l'on ajoute «$\wedge\neg M(a)$» à l'état-descriptif n° 1, la proposition résultante se contredit et peut donc être considérée comme donnant «trop» d'information.

La ressemblance la plus frappante a sûrement trait à la notion de *quantité d'information*; le lecteur aura sûrement d'ailleurs remarqué la similitude entre la définition de *inf* ci-dessus et celle de H chez Shannon. Il faut toutefois noter que, alors que Shannon ne s'intéressait qu'à la quantité d'information, Bar-Hillel s'emploie de plus à préciser la notion d'information en elle-même.

2.2.3 Théorie des situations

La question posée par Barwise est la suivante: Où se situe la signification d'un texte? Est-elle intrinsèque au texte, et donc complètement déterminée par celui-ci indépendamment de l'auteur ou du lecteur? Se situe-t-elle dans l'esprit de l'auteur ou du lecteur? Ou est-elle plutôt dans la vision du monde et des connaissances qui sont partagées par l'auteur et le lecteur?

La position de Barwise est de considérer la signification non pas uniquement du point de vue de l'énoncé, comme c'est le cas pour la théorie de l'information sémantique, mais aussi dans son contexte. Son approche consiste à séparer la signification littérale d'un énoncé de sa signification dans un contexte ou une situation donnée, qu'il appelle plutôt le *contenu*. La théorie résultante, connue sous le nom de *théorie des situations* [Bar89], est donc à la fois une théorie du sens et du contenu de l'information.

L'unité de base d'information dans la théorie des situations est l'*infon*. Les infons mettent en relation des objets en leur assignant une *polarité*, qui indique si la relation en question est vérifiée pour ces objets. Par exemple, dans l'infon:

$$\sigma_1 = \ll \acute{e}tudiant, A; 1 \gg$$

étudiant est une relation[4], A un individu, et 1 la polarité. Cet infon représente le fait que A est étudiant. La négation de σ_1 est obtenue en inversant sa polarité:

$$\sigma_1' = \ll \acute{e}tudiant, A; 0 \gg$$

L'infon σ_1 peut s'appliquer à différents énoncés. Imaginons par exemple trois individus A, B et C conversant entre eux. A dit à B: «*Je suis étudiant*». B dit à C: «*A est étudiant*». Enfin, C dit à A: «*Vous êtes étudiant*». Ces trois énoncés ont une signification différente, puisqu'ils sont distincts, mais leur contenu demeure le même – à savoir que A est étudiant – et est donné par σ_1.

La *composition* de l'information est donnée par la relation:

$$\sigma \oplus \sigma'.$$

Cette opération permet de fusionner ou d'unifier des informations compatibles entre elles. Si on pose $\sigma_2 = \ll \acute{e}tudiant, B; 1 \gg$, par exemple, alors:

$$\sigma_1 \oplus \sigma_2 = \ll \acute{e}tudiant, A \cup B; 1 \gg .$$

Il existe une relation de pré-ordre sur les infons donnée par:

$$\sigma \Rightarrow \sigma',$$

qui signifie que σ' est logiquement déduit de σ. Cette relation peut être déduite par les simples propriétés logiques de l'information. Par exemple, soit l'infon:

$$\sigma_3 = \ll \acute{e}tudiant, A, informatique; 1 \gg,$$

qui signifie qu'un individu A est étudiant en informatique. Alors $\sigma_3 \Rightarrow \sigma_1$, puisque le fait que A soit étudiant en informatique implique nécessairement le fait qu'il soit étudiant. La relation «$\Rightarrow$» peut aussi être donnée a priori. Elle exprime ainsi la *subsomption* entre infons. Par exemple, pour exprimer que les chênes sont des arbres, on dirait:

$$\ll ch\hat{e}ne, A; 1 \gg \Rightarrow \ll arbre, A; 1 \gg .$$

4. Même les relations peuvent être considérées comme des objets; ainsi, dans notre exemple, nous pourrions tout à fait «objectiviser» la relation pour parler d'«un étudiant».

Les infons peuvent être considérés dans le contexte d'*activités situées*, c'est-à-dire d'activités qui sont réalisées par des agents situés dans un environnement et qui affectent cet environnement. Il existe une relation liant les situations aux infons,

$$s \models \sigma,$$

qui est vérifiée pour une situation s et un infon σ, si ce dernier est valable dans la situation s (on dit aussi que s *supporte* σ). Supposant par exemple σ_1 tel que donné ci-dessus, et la situation s_1 où A prépare une thèse, alors $s_1 \models \sigma_1$ est vérifiée. Par contre, pour une autre situation s_2 où A est salarié, la relation $\models$ n'est pas vérifiée entre s_2 et A.

L'inclusion entre situations est définie comme suit: une situation s fait partie d'une situation s' si tout infon supporté par s l'est aussi par s'. En fait, cette relation d'inclusion définit un ordre partiel sur les situations.

Il est intéressant de noter que tout comme l'information sémantique n'établissait pas de lien entre valeur de vérité et information, de même, dans la théorie des situations, les infons et les situations ne sont pas rattachés à une valeur de vérité. En fait, Barwise oppose les notions de situations et d'infons à celle de *proposition*. Les propositions tirent leur valeur de vérité de l'association ou de la non-association d'un infon à une situation donnée. Ainsi, en reprenant l'exemple ci-dessus, la proposition $(s_1 \models \sigma_1)$ est $vraie$, mais la proposition $(s_2 \models \sigma_1)$ est $fausse$.

Cet exposé très succinct de la théorie des situations devrait suffire à démontrer son utilité. Nous n'avons pas abordé toutefois tout le problème de l'inférence, central pour Barwise puisqu'il valide sa théorie à l'aide de puzzles linguistiques et logiques. Ceci dépasse toutefois notre cadre de sens et d'information.

2.3 Information et langage

Nous regroupons dans cette section divers travaux de linguistique qui sont pertinents pour notre étude. Il ne s'agit pas d'une revue de littérature sur la linguistique, mais uniquement d'une investiguation des travaux qui peuvent nous aider à cerner la notion d'information. C'est pourquoi les problèmes classiques de traitement de langue naturelle, à savoir le choix du vocabulaire, la formation des énoncés, et l'interprétation de phénomènes linguistiques, ne sont pas traités ici. Pour les mêmes raisons, les aspects psycholinguistiques concernant la production et la reconnaissance du langage ne sont pas considérés.[5]

Les travaux portant sur l'identification et le rôle du thème dans la phrase et dans le texte sont présentés à la section 2.3.1. La section 2.3.2 concerne plutôt la signification du discours par la reconnaissance des intentions de l'énonciateur. Ces deux sections s'adressent principalement aux dimensions du signifié et de la pragmatique, respectivement. La dimen-

5. Voir [SDV95] à ce sujet.

sion du signifiant n'est pas reprise ici; elle pourrait inclure des travaux de morphologie, d'étymologie ou de sémiotique.

2.3.1 Thème

a) Analyse statutaire

L'*analyse statutaire*, telle qu'introduite par Zemb [Mel87], cherche à retrouver le statut logique et les relations entre les éléments du langage. Elle s'inscrit comme l'une des quatre dimensions possibles d'analyse du langage, les trois autres étant: la *sémiotique*, c'est-à-dire l'étude des signes; la *rhétorique*, ou l'étude du discours; et enfin, le *superficiel*, ou la grammaire de surface. Ces dimensions peuvent varier indépendamment, ce qui veut dire notamment que la structure statutaire ne dépend pas de la structure de surface de la langue.

Au premier niveau[6] de l'analyse statutaire, chaque élément de la phrase a l'un de ces trois status: il est soit *thème*, *phème*, ou *rhème*.[7] Selon Zemb, toute communication s'effectue par l'application du rhème au thème, par le phème. Le thème affirme que quelque chose existe dans l'univers du discours, le rhème affirme quelque chose à propos du thème, et le phème lie les deux ensembles. Plus intuitivement, le thème correspond au sens courant, c'est-à-dire à «ce dont on parle», le rhème à «ce qu'on en dit», et le phème, à «comment on le dit».

Plusieurs exemples sont donnés ci-dessous. Mais d'abord, puisqu'en pratique le thème et le rhème peuvent être assez difficiles à identifier dans une phrase, les règles suivantes sont définies:

- *test temporel.* Le thème d'une proposition inclut toujours une indication du temps, le plus souvent donnée implicitement par la conjugaison du verbe (le temps sera alors présent, passé, futur, etc.). Si un élément de la phrase est une spécification de temps explicite, alors cet élément fait partie du thème. Par exemple: «*aujourd'hui* », «*la semaine dernière* », «*bientôt* », etc.

- *test d'attribut.* Si un complément d'objet direct et son verbe peuvent être paraphrasés par un verbe alternatif, alors ils font partie du rhème. Par exemple: «*il a perdu du sang*» devient «*il a saigné*», «*il a pris des mesures (métriques)*» devient «*il a mesuré*».

- *test du nom-prédicat.* Si un complément d'objet direct et son verbe peuvent être paraphrasés comme un *nom-prédicat*, c'est-à- dire un groupe nominal où le prédicat a été

6. Le second niveau, que nous ne verrons pas ici, vise l'assignation d'une structure interne aux éléments identifiés au premier niveau.

7. «thème» est ici la traduction de Zemb du grec «onoma», qui a traditionnellement été assimilé à «sujet»; «rhème» vient aussi du grec; quant à «phème», il est dû à Charles S. Peirce

«nominalisé», alors ils font partie du rhème. Par exemple, «*il vend des ordinateurs*» devient «*c'est un vendeur d'ordinateurs*».

- *test d'intégration.* Ce test procède en prenant un élément du thème connu et en tentant d'y intégrer un second élément, afin de déterminer si ce dernier fait aussi partie du thème. Supposons par exemple «*Le tramway est en service depuis peu à Strasbourg*». «*Le tramway*» est dans le thème. Pour déterminer si «*à Strasbourg*» fait aussi partie du thème, on peut paraphraser comme suit: «*Le tramway de Strasbourg est en service depuis peu*». Donc «*à Strasbourg*» fait aussi partie du thème.

- *test de négation.* Ce test ne s'applique qu'aux phrases affirmatives. Il s'agit de transformer la phrase affirmative en négative, et d'examiner ce qui demeure positif, ou dont l'existence est toujours affirmée, et ce qui devient rejeté, c'est-à-dire qui est nié. Le positif appartient au thème, alors que le négatif appartient au rhème. Par exemple, soit «*Le tramway est en service à Strasbourg*». Dans la négation de cet énoncé, «*Le tramway n'est pas en service à Strasbourg*», «*tramway*» et «*Strasbourg*» existent toujours, et font donc partie du thème; par contre l'acte d'être «*en service*» est nié et appartient donc au rhème.

- Il peut sembler à première vue que le complément d'objet direct fait partie du thème quand il est introduit par un article défini, et du rhème sinon. En effet, l'article défini s'emploie généralement pour parler de quelque chose de particulier, et donc dont l'existence est affirmée. Malheureusement il existe des contre-exemples pour les deux cas. Dans l'expression «*il a le hocquet*», «*le hocquet*» fait partie du rhème et non pas du thème, puisque sa négation, «*il n'a pas le hocquet*», nie l'existence du hocquet (pas en tant que concept générique bien entendu, mais uniquement dans le contexte de «*il*»!). De même dans «*il cherche un livre qu'il a perdu la semaine dernière*», on voit bien que «*un livre*» réfère à un livre particulier, et donc appartient au thème. Si, pour l'anglais, l'article «a» est une bonne indication du rhème, l'article «the» et d'autres articles indéfinis ne sont pas des indices fiables. En français, l'ambiguïté sur «un», qui est tantôt article indéfini, adjectif numéral, pronom indéfini, etc., complique encore davantage la situation.

Voyons maintenant l'application de ces règles à l'aide de quelques exemples.

(2.1) *Le vendredi il mange à la cafétéria du campus.*

«*Le vendredi*» est un marqueur de temps et appartient donc au thème (*test temporel*). Appliquons le test de négation pour obtenir «*Le vendredi il ne mange pas à la cafétéria du campus*». «*Il*» existe toujours bien sûr, alors que le prédicat «*mange à la cafétéria du campus*» est nié. Reste à savoir si «*la cafétéria du campus*» fait partie du thème ou du rhème. Elle appartient plutôt au thème puisque la cafétéria existe toujours dans la négation de l'énoncé. Finalement, le thème est donc formé de «*le vendredi*», «*il*», «*à la cafétéria du*

campus», et du marqueur *présent*. Le rhème ne contient que «*mange*». Le phème, enfin, est donné par les marqueurs *affirmatif* et *indicatif*.

(2.2) *Il mange du poisson.*

Dans la négation de cet énoncé, «*Il ne mange pas de poisson*», le poisson n'existe pas nécessairement. On en déduit donc que le rhème est donné par «*mange du poisson*». Le thème, quant à lui, est donné par «*il*», *présent*. Le même résultat est obtenu par le test du nom-prédicat, en paraphrasant l'énoncé par «*Il est un mangeur de poisson*».

L'analyse statutaire telle que proposée par Zemb est indépendante de formulations particulières telles que l'emploi de la forme passive ou active. Nous avons déjà vu avec les exemples ci-dessus des thèmes qui étaient sujet ou complément. Le rhème peut aussi être sujet ou prédicat. Ainsi, dans l'énoncé suivant:

(2.3) *Un livre est posé sur la table.*

«*un livre*» est à la fois rhème et sujet. On peut s'en convaincre en appliquant une fois de plus le *test de négation*; dans «*un livre n'est pas posé sur la table*», aucune existence n'est prêtée à «*un livre*».

Une conclusion est que la traditionnelle dichotomie entre *sujet* et *prédicat*, telle qu'enseignée par les grammaires modernes, est inappropriée pour représenter la structure logique d'une phrase, puisqu'elle est dépendante de la structure superficielle du langage, ce qui n'est pas le cas de l'analyse statutaire.

b) Progression thématique

Si les travaux d'analyse statutaire présentés ci-dessus examinent la structure logique de la phrase, ils ne peuvent expliquer la structure inter-phrase, qui obéit à des principes complètement différents. Vue dans son contexte, la phrase ne peut pas être comprise uniquement de façon statique, mais doit être prise comme une série d'instructions donnée par l'auteur au lecteur. Afin de reconnaître ces instructions, et donc de comprendre la phrase, le lecteur dispose d'indices dans la phrase lui permettant de retrouver les éléments dits «connus» de ceux qui sont «nouveaux».

Cette vision de la structure logique du texte est appelée TFA (*Topic-Focus Articulation*) ou *progression thématique* [HS84]. Les éléments d'une phrase sont divisés entre le *sujet* (*topic*)[8] et le *commentaire* ou *centre d'intérêt* (*focus*). Le sujet correspond aux éléments d'information qui sont partagés entre l'auteur et le lecteur, et que l'auteur considère

8. Le lecteur aura remarqué que «thema» (analyse statutaire) a été traduit par «*thème*», alors que le «topic» (TFA) a été traduit par «*sujet*». Il est montré à la section 2.5 comment ces deux notions sont effectivement reliées.

facilement accessibles pour le lecteur à ce point dans le discours. Le commentaire, lui, désigne les propriétés ou modifications apportées à ce sujet, ou de nouvelles relations avec d'autres éléments du discours.

Le commentaire est souvent mis en relief dans la phrase, soit par des procédés syntaxiques dans le discours écrit ou par l'emploi d'intonations dans le discours parlé. Ainsi, dans l'énoncé 2.4a, la réponse à la question sera différente selon que le locuteur aura mis l'accent sur «*ce train*» ou sur «*tous les jours*».[9] Dans 2.4b, par contre, l'ambiguïté est levée par la structure syntaxique.

(2.4) *(a) Prend-il ce train tous les jours?*
(b) Est-ce ce train qu'il prend tous les jours?

En l'absence de telles informations, le contexte doit pouvoir permettre de trancher. Ainsi, la phrase 2.5, ne permet pas en elle-même de déterminer le rôle de «*la main*» et de «*avec le couteau*».[10] Par contre, si l'on sait qu'il s'agit de la réponse à la question «*Comment s'est-il coupé la main?*», alors «*avec le couteau*» fait forcément partie du commentaire.

(2.5) *Il s'est coupé la main avec le couteau.*

Sujet et commentaire ne sont pas sans rapport avec les notions d'éléments *liés* ou *non-liés contextuellement*. Les éléments liés sont ceux qui font référence à un élément apparaissant dans une partie précédente du discours, alors que les éléments non-liés sont ceux qui apparaissent pour la première fois dans le discours.

Le sujet et le commentaire sont aussi souvent assimilés au paradigme d'information *donnée* (*given*) et information *nouvelle* (*new*). L'information donnée est celle que le locuteur suppose connue de l'auditeur, alors que l'information nouvelle est ce qu'il veut lui apporter de nouveau. Une mise en garde contre une association trop rapide entre ces notions est donnée dans [HS84]. En effet, le sujet n'est pas nécessairement supposé connu par le locuteur, comme c'est le cas avec l'information donnée. Il suffit qu'il soit lié contextuellement; son existence peut même être niée.

Le test de la question peut être utilisé afin de déterminer la structure TFA d'une phrase. Il s'agit de trouver l'ensemble $\mathcal{Q}$ des questions qui peuvent être répondues directement et complètement par la phrase, puis d'utiliser une de ces règles:[11]

– **Sujet**. Si A apparaît dans tous les éléments de $\mathcal{Q}$, il est alors *sujet*;

9. Si cet emploi de l'intonation est plutôt rare en français, où l'intonation suit plutôt des règles de *rythme*, il est obligatoire dans d'autres langues telles que l'anglais, où l'intonation est liée de près au *sens*.

10. Du moins en français, puisqu'en anglais l'un de ces deux éléments recevrait nécessairement l'accent de phrase.

11. Seules les deux règles les plus simples sont données; la liste complète peut être trouvée dans [HS84].

– **Commentaire**. Si A (de la phrase à tester) n'apparaît dans aucun élément de $\mathcal{Q}$, il est alors *commentaire*.

Par exemple, pour l'énoncé 2.5, l'ensemble des questions $\mathcal{Q}$ comprend: «*Que lui est-il arrivé?*», «*Comment s'est-il coupé la main?*», «*Que s'est-il coupé avec le couteau?*», etc. «*il*» est pré-supposé dans tous ces contextes, et fait donc partie du sujet de 2.5.

Jusqu'à présent nous avons vu que le sujet et le commentaire s'inscrivaient bien dans un contexte particulier, mais ne les avons déterminés que pour des phrases isolées. Il est intéressant de voir comment ces éléments se coordonnent au niveau inter-phrase, comme nous l'avions d'ailleurs suggéré au début de cette section. Dans [Dan74], une explication de la progression du sujet entre les phrases est donnée.

Le premier type de progression dans un texte se produit quand le commentaire d'une phrase S_1 réapparaît dans la phrase suivante S_2. Dans l'exemple ci-dessous, «*Prague*» est dans le commentaire de la première phrase et dans le sujet de la seconde.

(2.6) *Cet été j'ai visité Prague.*
C'est une ville magnifique.

Le second type se présente quand le sujet de S_1 est identique au sujet de S_2, ou si l'un de ces deux sujet inclut l'autre. Les phrases suivantes ont toutes deux «*Prague*» pour sujet:

(2.7) *Prague est connue comme «la ville aux cent clochers».*
Elle possède aussi plusieurs ponts dont l'architecture est remarquable.

Enfin, dans le troisième type, les deux phrases S_1 et S_2 dérivent leur sujet d'un *hyper*-sujet commun. Ainsi, les deux phrases ci-dessous ou pour hyper-sujet «*Prague*».

(2.8) *Le pont Charles est très prisé des visiteurs.*
La ruelle d'Or, à l'origine appelée ruelle des Orfèvres, est aussi très visitée.

2.3.2 Intentions

a) Maximes de conversation

Toute recherche de sémantique linguistique, dès qu'elle s'approfondit quelque peu, tend à faire intervenir des déterminations d'ordre pragmatique. Situé dans un contexte de communication, un énoncé ne prend sa signification que si l'agent à qui il est destiné est en mesure de le comprendre. C'est dans cette optique que Grice définit la signification d'un énoncé:

> «l'intention du locuteur que son énonciation produise un certain effet sur un auditoire, au moyen de la reconnaissance de cette intention» [Gri71]

Pour que cet effet soit atteint, la communication doit suivre un principe général, que Grice appelle le *principe de coopération*, et qui stipule que l'énonciateur doit se conformer aux quatre règles suivantes: [Gri75]

1° *Règle de la quantité.* L'énoncé doit contenir autant, mais pas plus, d'information qu'il n'est requis.

2° *Règle de la qualité.* L'énoncé doit être véridique, du moins du point de vue de l'énonciateur.

3° *Règle de relation.* L'énoncé doit être pertinent.

4° *Règle de modalité.* L'énoncé doit être clair, sans ambiguïté.

Une critique de ces travaux est présentée dans [Car83]. Il est clair qu'il ne s'agit pas de règles normatives; ainsi, la maxime de la *qualité* ne signifie pas que tout énonciateur ne dise que ce qu'il croit être vrai, mais plutôt qu'il se *présente* comme tel. C'est ce qui se produit dans le cas d'un mensonge, par exemple.

b) Structure du discours

Les travaux de Grosz et Sidner sur la structure du discours [GS86] reprennent cette idée d'*intention* et en font un élément central dans leur étude du discours.

Le *discours* est à prendre ici comme une manifestation du langage qui comprend plusieurs énoncés et implique typiquement plusieurs participants. Un de ces participants *initie* le discours, et est appelé *ICP* (*initiating conversational participant*). Les autres participants, dénommés par *OCP* (*other conversational participant*), sont initialement auditeurs, mais peuvent par la suite être eux-mêmes énonciateurs. Notons que même dans un texte, où la communication se résume à un monologue de l'ICP, l'OCP joue un rôle. En effet, le texte lui étant destiné, l'ICP lui adresse donc indirectement la parole.

De la même façon que les mots se combinent pour former des groupements, qui à leur tour forment des phrases, les énoncés se combinent entre eux pour former des *segments de discours* (*DS*), qui forment à leur tour le discours entier. Dans un texte, ces segments de discours sont souvent délimités par des changements de paragraphes, sections, etc., ou par l'emploi de certains mots ou expressions. Cette factorisation en segments de discours n'est pas exclusive au discours écrit, mais a été observée pour une variété de types de discours.

Une des propriétés essentielles d'un discours ou d'un segment de discours est qu'il a toujours un *but* particulier. De façon intuitive, ce but correspond à l'*intention* de l'énonciateur: il peut s'agir de la raison pour laquelle le discours a été engagé, ou de la raison pour laquelle une information plutôt qu'une autre est incluse dans son discours. Cependant, seules les intentions qui sont destinées à être reconnues par l'auditeur peuvent être considérées comme des *buts*. Les intentions privées de l'énonciateur, comme par exemple celle d'enseigner ou d'impressioner l'interlocuteur à son insu, ne sont donc pas considérées.

Les buts d'un discours ou d'un segment de discours sont désignés par *DP* (*discourse purpose*) et *DSP* (*discourse segment purpose*), respectivement.

Supposons la situation suivante, où deux «agents», Jérémie et Iris, doivent acheter des billets d'avion. Jérémie s'adresse à Iris par l'une des phrases suivantes:

(2.9) *(i) Peux-tu t'occuper des billets?*
(ii) Je les ai achetés cet après-midi.
(iii) Je suis passé à l'agence cet après-midi.

Voici quelques exemples de DSP pour cette situation:

- l'intention qu'un agent effectue une tâche. Par exemple, l'intention de Jérémie que «Iris achète les billets d'avion», dans l'énoncé 2.9i;
- l'intention qu'un agent croit un fait. Par exemple, l'intention que «Iris croit que j'ai acheté les billets d'avion», dans l'énoncé 2.9ii;
- l'intention qu'un agent croit qu'un fait en supporte un autre. Par exemple, l'intention que «Iris croit que si je reviens de l'agence de voyage, c'est que j'ai acheté les billets d'avion», dans l'énoncé 2.9iii.

Il est clair qu'aucune liste de ce genre ne saurait être exhaustive; les types de DP/DSP sont aussi variés que les types de personnes qui peuvent prendre part à un discours, et les types de situations qui peuvent se présenter à elles.

Deux relations structurelles sont définies entre les DP/DSP: la *domination* et la *précédence*. On dit que le DSP_1 *domine* le DSP_2, ou de façon équivalente que le DSP_2 *contribue* au DSP_1, si le DSP_2 est destiné à satisfaire en partie le DSP_1. Par exemple, le but de cette sous-section 2.3.2 est d'exposer quelques travaux ayant trait à l'information et aux intentions dans un discours. Ce but participe à celui plus général de la section 2.3, qui est de présenter les travaux ayant trait à l'information et à la linguistique. On dit donc que $DSP_{\S 2.3}$ domine $DSP_{\S 2.3.2}$, noté par $DSP_{\S 2.3}$ $\mathcal{D}om$ $DSP_{\S 2.3.2}$.

La relation de précédence quant à elle, est surtout utile dans les discours où l'aspect temporel est important. On dira que DSP_1 *précède* DSP_2 si DSP_1 doit être satisfait avant DSP_2. C'est le cas entre autres quand le discours consiste en des instructions que l'ICP donne à l'OCP afin d'effectuer une tâche.

Imaginons un texte D qui tente de rétablir des idées fausses à propos du Père Noël, et supposons les deux phrases données en 2.10 en guise d'introduction à ce texte. Chacune de ces phrases, puisqu'elle énonce des idées différentes, correspond à un segment de discours. Appelons DS_i le segment correspondant à la phrase i, et DS_{ii} le segment pour la phrase ii. L'intention reliée à la phrase DS_i, le DSP_i, peut s'énoncer comme suit: l'intention de l'ICP (l'auteur du texte) de convaincre l'OCP (le lecteur) que personne ne peut croire «toutes les absurdités dites à propos du Père Noël». Le DSP_{ii}, lui s'énonce comme suit: l'intention

de l'ICP de convaincre l'OCP qu'il est grand temps de rectifier les faits pour le bénéfice des générations futures. On a donc $D\ \mathcal{D}om\ DSP_i$ et $D\ \mathcal{D}om\ DSP_{ii}$.

(2.10) *i) Qui pourrait croire toutes les absurdités qui ont été dites à propos du Père Noël?*
ii) Il est grand temps de rectifier les faits pour le bénéfice des générations futures.

Les DSP dans cet exemple sont des intentions de l'ICP de convaincre l'OCP de quelque fait, comme c'est souvent le cas pour des textes descriptifs.

Un point intéressant pour notre étude est l'équivalence qui est démontrée dans [GS86] entre le *thème* (*topic*) du discours et le DP ou le DSP d'un segment. Au niveau des phrases, Grosz et Sidner préfèrent parler de *centre*, qui correspond à l'élément central d'un énoncé. Cependant la correspondance entre le thème d'une phrase et le centre d'un énoncé est imparfaite.

Les intentions qui servent de DP/DSP sont une extension naturelle aux intentions de Grice. Elles se situent à un niveau plus général que d'autres analyses de rhétorique [MMT89] ou de cohésion lexicale [HH76] [MH91], aussi sont-elles indépendantes de la langue ou du type de discours. Cependant, elles ne peuvent tenir compte à elles seules de la signification d'un discours.

Notons enfin que des relations d'un autre ordre participent également à la structure d'un discours. Grosz et Sidner considèrent les structures linguistiques et les structures d'*attention*, ces dernières étant une abstraction de la progression du centre d'intérêt à mesure que le discours se déroule.

2.4 Information et Recherche d'Information

Nous avons vu jusqu'à maintenant des théories provenant de diverses domaines qui peuvent fournir des hypothèses quant à la notion d'information. De telles hypothèses sont malheureusement absentes des modèles de recherche d'informations, aussi devons-nous adopter ici une approche plus pragmatique, et examiner plutôt quelle interprétation a été donnée à cette notion jusqu'à maintenant dans les systèmes de recherche d'informations.

Nous tentons pour ce faire de classifier les systèmes de recherche d'informations selon leur définition et leur emploi de l'information, et les évaluons sur ces critères. Nous abordons ensuite trois aspects spécifiques à la recherche d'informations, qui devraient également être considérés dans un modèle d'indexation: la pertinence de l'information, le besoin de l'utilisateur, et la structure des documents.

2.4.1 L'information dans les systèmes existants

Dans cette section, nous examinons comment l'information a été représentée et traitée jusqu'à maintenant dans les systèmes de recherche d'informations. Pour ce faire, il est nécessaire d'affiner la classification signifiant-signifié-pragmatique, utilisée jusqu'à présent, en ne considérant plus uniquement le niveau de représentation, mais aussi le niveau d'analyse pour extraire cette représentation. Nous classons donc les systèmes selon deux axes: i) suivant la complexité de représentation de l'information qu'ils utilisent, et ii) selon le processus d'indexation qu'ils emploient. D'autres axes, comme l'inférence ou la mise en correspondance des informations, pourraient aussi être définis, mais dépassent le cadre de notre intérêt pour un modèle d'indexation.

L'axe de représentation de l'information correspond à la complexité des index représentés par les systèmes. Ces catégories ne font pas directement référence aux aspects signifiants ou signifié de l'information: en effet, ces deux notions sont souvent confondues en recherche d'informations. Si leur distinction peut s'avérer utile dans une étude linguistique, elle est bien souvent superflue en recherche d'informations: le but n'étant pas de comprendre les documents, mais uniquement de permettre une mesure de similarité avec une requête.

Au niveau le plus bas de la représentation de l'information, on retrouve l'indexation traditionnelle par *mots-clés*. La catégorie *groupes* réfère aux systèmes qui groupent les mots-clés simples pour former des termes complexes, généralement sur la base de critères syntaxiques.[12] La catégorie *surface* réfère à des structures dites «de surface» du texte, où sont représentées les structures syntaxiques de phrases ou de groupes nominaux. Enfin, la catégorie *concept* englobe les systèmes qui représentent le sens des documents, c'est-à-dire les concepts et les relations entre ces concepts. On parle souvent à ce niveau de recherche d'informations conceptuelle (*conceptual information retrieval*).

L'échelle d'*indexation* est essentiellement basée sur l'analyse linguistique qui est réalisée par le système. Les systèmes dans la catégorie *manuelle* n'effectuent pas de telles analyses. L'indexation par *extraction* réfère non pas à une analyse linguistique, mais à un processus de sélection des concepts importants dans un texte, par la reconnaissance de certaines expressions ou d'indices linguistiques.[13] La *lemmatisation* consiste à normaliser les mots par le retrait de suffixes. L'analyse syntaxique cherche le groupement de ces mots en des structures syntaxiques. Enfin, l'analyse *sémantique*, on l'aura deviné, vise à déduire automatiquement le sens des documents.

Une analyse pragmatique, que nous ne considérons pas ici, est aussi possible. Peu de travaux en recherche d'informations font directement intervenir les aspects pragmatiques. Citons toutefois [BH94], qui est une application directe de la théorie des situations à la recherche d'informations, ainsi que [Hea94], où une application des théories du discours est tentée par la segmentation multi-paragraphe.

12. Cette stratégie est appelée *phrase indexing* dans la littérature.
13. On réfère aussi à ce processus par *text skimming* dans la littérature.

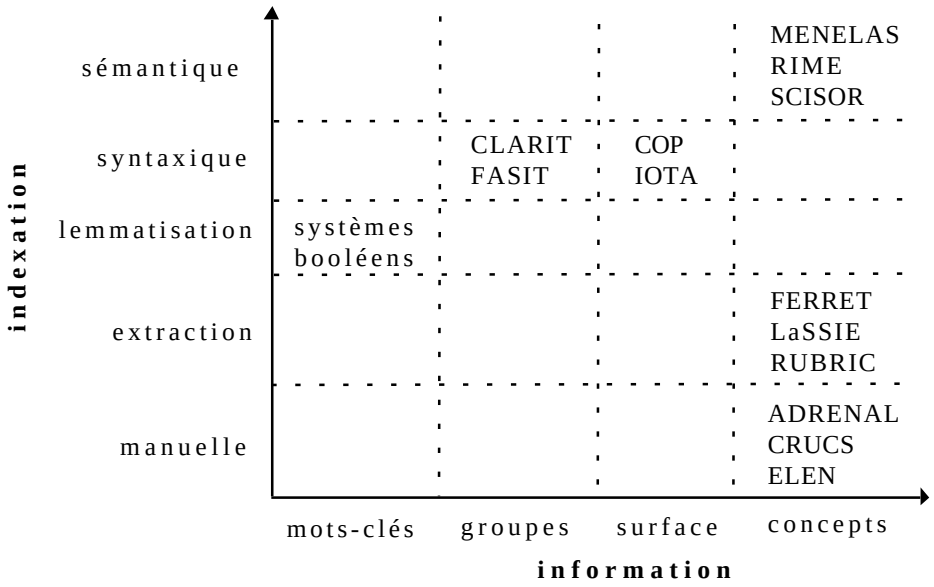


Figure 2.2. Classification des systèmes de recherche d'informations

La figure 2.2 positionne 14 systèmes de recherche d'informations parmi les plus connus ou les plus représentatifs de leur catégorie selon cette classification. Il va sans dire qu'une telle classification doit être prise sous toute réserve, puisque la frontière entre les catégories n'est pas toujours claire. Quand un système pourrait appartenir à plusieurs de ces catégories, nous l'avons placé sous sa catégorie dominante, ou la catégorie qui correspond au point fort de ce système. Les points plus élevés sur les axes sont privilégiés: ainsi, à défaut d'une composante dominante, un système incorporant des analyses syntaxiques et sémantiques sera placé sous la catégorie *sémantique*. Cela ne présume en rien toutefois de l'efficacité et de la performance de ces systèmes: il est tout à fait possible qu'un système ayant une représentation de l'information très primitive produise de meilleurs résultats qu'un autre ayant une représentation plus poussée.

Nous examinons maintenant plus en détails ces classes, et offrons ensuite une évaluation sommaire .

a) Lemmatisation & mots-clés

Les archétypes de cette classe sont les systèmes booléens traditionnels et le système vectoriel [SM83], encore largement les plus répandus dans l'industrie. Les mots-clés utilisés par ces systèmes se situent à mi-chemin entre le signifiant et le signifié. Ils relèvent bien sûr avant tout de la structure superficielle linguistique, dont ils sont issus, mais la lemmatisation les rapproche d'un niveau conceptuel, puisqu'elle cherche à extraire la *racine* commune à

un ensemble de mots. Par exemple, «*information*», «*informateur*» et «*informer*» ont comme racine commune, «*inform*». Il ne saurait s'agir d'un niveau purement conceptuel, toutefois, puisque cette racine n'est pas désambiguïsée. Si elle ne correspond pas encore à un concept, la racine peut s'en rapprocher davantage par l'emploi d'un thésaurus.

b) Analyse syntaxique & groupes

Pour pallier à la représentation trop pauvre des mots-clés, une solution consiste à utiliser des groupes de mots-clés comme index. Bien que la méthode de prédilection pour identifier ces groupes soit une analyse syntaxique, on peut aussi procéder par regroupements statistiques, en observant les fréquences de co-occurences des mots-clés [Fag88][LC90]. Lorsqu'une analyse syntaxique est utilisée, les groupes correspondent alors à des unités syntaxiques de taille fixe ou variable: par exemple, aux groupes nominaux.

Dans le système CLARIT [Pai93], les groupes sont obtenus après une analyse syntaxique robuste qui identifie les groupes nominaux dans les textes, et qui en extrait des groupes candidats. Après avoir calculé un poids pour chacun de ces groupes, ils sont ensuite classifiés en trois catégories: les termes exacts, généraux, et nouveaux. Pour ce faire on utilise un ou des dictionnaires de termes certifiés; si une correspondance parfaite est trouvée, le terme est dit exact, s'il est une partie ou un sous-terme d'un terme certifié, il est général, sinon, il est nouveau. La figure 2.3 donne un exemple de représentation dans CLARIT, avec la classification des groupes selon les trois catégories. Les nombres à gauche des groupes correspondent aux poids calculés. CLARIT a été testé sur plusieurs corpus provenant de domaines très variés, citons entre autres une étude récente sur un corpus d'articles du Wall Street Journal dans [Pai93].

termes exacts		*termes généraux*		*termes nouveaux*	
2.32	nasdaq	5.39	pretroleum	2.81	prnewswire giant
1.16	est	5.39	prnewswire	2.69	prnewswire
0.38	toronto	0.92	asset	0.12	giant
0.37	wholly owned			1.98	pacific petroleum
	subsidiary				incorporated

Figure 2.3. Représentation CLARIT d'un document (tiré de [Pai93, p.386])

A l'instar de CLARIT, le système FASIT [DG83] (*Fully Automatic Syntactic Indexing of Text*) tente lui aussi de pallier les inconvénients de l'indexation par termes en regroupant les termes. Chaque texte est décomposé mot par mot en assignant tout d'abord à chaque mot une catégorie syntaxique de manière à extraire des termes ou des syntagmes supposés forts selon des critères syntaxiques consignés dans des schémas syntaxiques prédéfinis. Des

expériences ultérieures ont proposé une analyse syntaxique plus poussée [BD92] utilisant des réseaux de transitions récursifs (*recursive transition networks*). Le corpus utilisé dans cette expérience consistait en des références d'articles de MEDLINE (*National Library of Medicine*) entre 1974 et 1979 concernant la fibrose kystique.

c) Analyse syntaxique & structure de surface

COP (Constituent Object Parser) [MHCW89] réalise aussi une analyse syntaxique, mais cette fois en utilisant une représentation plus poussée, basée sur les dépendances syntaxiques. L'accent est mis sur la structure hiérarchique des entités: plutôt que d'essayer de représenter le détail fin, on se contente de représenter les propriétés structurelles des descriptions.

L'analyse syntaxique est réalisée par une grammaire robuste, qui sur-accepte le langage, et qui produit autant que possible des représentations canoniques. Bien que cette grammaire accepte un large sous-ensemble de la langue naturelle, il n'en demeure pas moins qu'elle a été conçue pour un type de corpus particulier (des résumés d'articles): ceci permet entre autre l'application d'heuristiques bien spécifiques pour résoudre les problèmes d'ambiguïtés, ellipses, anaphores, etc.

Les structures produites sont des arbres binaires, où les arcs indiquent si deux constituants sont liés, et une étoile indique le constituant dominant. La figure 2.4 donne un exemple de représentation dans COP. La flèche dans cet exemple souligne une référence anaphorique entre «*cop*» et «*who*».

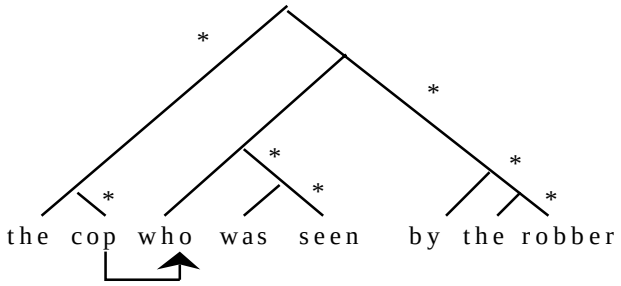


Figure 2.4. Exemple de représentation dans COP (tiré de [MHCW89])

Le système IOTA [CD87] est aussi à placer dans cette catégorie. L'analyse morpho-lexicale est très poussée dans IOTA, allant même jusqu'à permettre l'enrichissement automatique du vocabulaire lors de la rencontre de mots nouveaux [Pal90]. L'indexation est réalisée de façon automatique par l'extraction de groupes nominaux, qui sont ensuite utilisés pour la génération de termes d'indexation. L'indexation s'aide d'un thésaurus qui est construit automatiquement par des critères statistiques et linguistiques, et tient compte de

la structure du texte[14]. L'interrogation utilise également le thésaurus et adapte les réponses selon le profil de l'utilisateur.

d) Indexation manuelle & concepts

Nous entendons par *concept* une représentation sémantique indépendante de la langue, qui peut se limiter aux concepts simples, mais qui comprend aussi généralement des liens ou relations entre concepts. La représentation de ces concepts prend diverses formes: *frames*, graphes, formules logiques, etc.

ADRENAL (*Augmented Document Retrieval using Natural Language*) [CL87] est construit au-dessus d'un système également développé par l'équipe de Croft: I^{3}R. L'accent dans ce système est mis sur la représentation des documents et des requêtes dans un formalisme adéquat. Dans ce but, les auteurs ont développé REST (*Representation for Science and Technology*), qui est un langage basé sur les *frames* et pour lequel ils ont défini quelques concepts scientifiques de base. La figure 2.5 donne un exemple de représentation dans ce formalisme. Chaque *frame* est associé à un type correspondant à un concept tel que STUDY, METHOD, etc. Le *slot* APPEARANCE identifie quel mot du texte correspond à ce concept. Les autres *slot* définissent les liens entre concepts.

```
(STUDY-1                                (METHOD-1
    APPEARANCE: analysis)                   APPEARANCE: algorithm)

(STUDY-2                                (METHOD-2
    IS-A: STUDY-1                           IS-A: METHOD-1)
    ARGUMENT-OF: RELATIONSHIP-1)            USES: ACTION-1

(STUDY-3                                (METHOD-3
    IS-A: STUDY-2                           IS-A: METHOD-2)
    INTEREST: METHOD-3)                     APPEARANCE: Quicksort

(RELATIONSHIP-1                         (ACTION-1
    APPEARANCE: probabilistic)              APPEARANCE: divide and conquer)
```

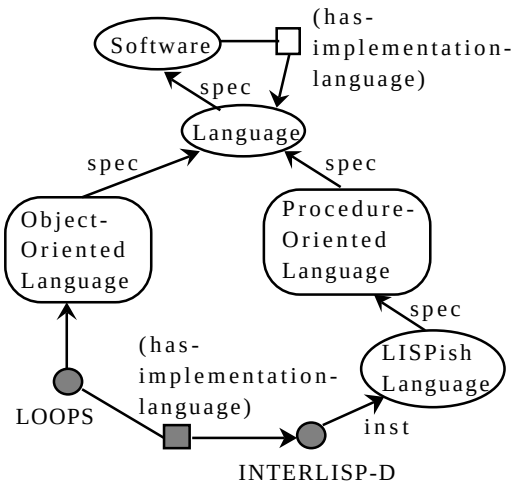
Figure 2.5. Représentation ADRENAL de «*Probabilistic analyses of Quicksort, a divide and conquer algorithm*» (tiré de [CL87, p.27])

Le processus de recherche dans ADRENAL s'effectue d'abord par un pré-filtrage des documents à l'aide de mots-clés. Les documents ainsi sélectionnés peuvent alors être analysés

14. Cet aspect sera abordé plus en détails à la section 2.4.3c)

afin de trouver leur représentation REST, et de les comparer avec la requête. Malheureusement, certains composants du système, telle que l'analyse automatique des textes en langue naturelle, semblent n'avoir jamais été réalisés. Ainsi, les résultats présentés sur le corpus CACM dans [CL87] ont été obtenus après indexation manuelle.

CRUCS (*Conceptual Retrieval Using Connectionist Style*) [BM88] est un prototype pour utiliser la représentation de connaissances et l'inférence en recherche d'informations. Il utilise pour ce faire un langage de représentation de la connaissance, μKL-ONE, qui comme on s'en doute est de la famille KL-ONE. Ce langage permet la modélisation des documents et des connaissances par une taxonomie de concepts. La figure 2.6a donne un exemple de hiérarchie de concepts dans CRUCS. Dans cet exemple, *LOOPS* et *INTERLISP-D* sont des instances des concepts *Object-Oriented-Language* et *Procedure-Oriented-Language*, eux-mêmes sous-classes de *Language*.



a) taxonomie des concepts

sujet: object-oriented languages
unités: Software, Language,
Object-Orientedness
rôle: has-implementation-language
unités: has-part,
has-implementation-language
domaine: Software; **range**:Language
restriction: INTERLISPish-Language
unités: Sofware, Language,
Procedure-Oriented-Language,
LISPish-Language,
INTERLISPish-Language

b) requête « *object-oriented language implemented in an 'INTERLISPish' language*

Figure 2.6. Exemple de représentation dans CRUCS (tiré de [BM88])

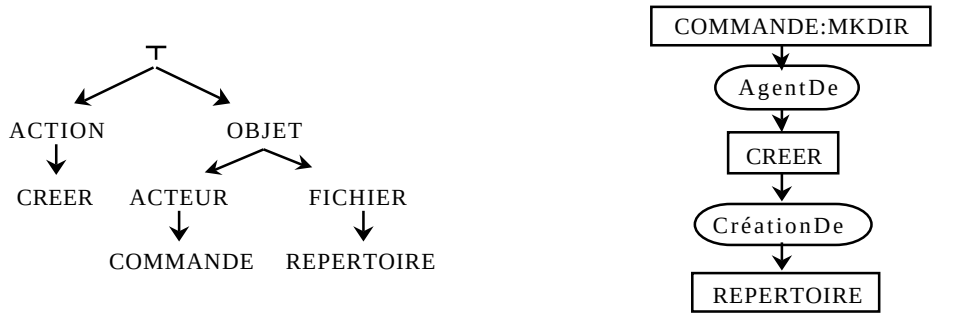
L'originalité de ce système réside dans l'approche connectionniste qu'il utilise pour la recherche. A partir des structures symboliques et hautement hiérarchisées qui sont utilisées pour modéliser les documents et le domaine d'application, une correspondance vers un niveau non-symbolique et non-structuré est réalisée par le biais de micro-unités. Ces unités sont des éléments atomiques que les concepts du domaine peuvent avoir en commun. Il peut s'agir de concepts primitifs, de restriction de nombre, de valeur, etc. Chaque concept est alors représenté par un ensemble de liens valués vers ces micro-unités, qu'il peut aussi hériter de ses concepts supérieurs. Les requêtes sont représentées de la même manière, sauf qu'elles sont divisées en trois parties: le sujet, le rôle, et une restriction sur le rôle. La

recherche se fait alors par inférence dans un réseau de neurones.

CRUCS a été testé sur un corpus portant sur des descriptions d'outils pour l'Intelligence Artificielle, provenant du guide de Waterman «*A Guide to Expert Systems*» et d'autres descriptions créées pour le besoin de l'expérimentation.

ELEN (*géniE logicieL & recherchE d'informatioNs*) [Che92] est un prototype de système de recherche d'informations orienté vers la précision des réponses. Il utilise comme formalisme de représentation les graphes conceptuels. Les documents comme les requêtes sont traduits dans ce langage, mais ces dernières sont d'abord entrées par l'utilisateur dans un pseudo-langage graphique qui guide la formation des requêtes, puis traduites par le système en graphes conceptuels. La figure 2.7 donne un exemple de représentation dans ELEN. En a) on voit le treillis de types, similaire bien que moins expressif que la taxonomie de concepts dans CRUCS. En b) un document est montré: les rectangles correspondent aux concepts et les ovales, aux relations entre concepts.

La recherche dans ELEN se fait par la *projection* d'une requête sur les documents. Afin d'optimiser cette opération, on fait une pré-sélection des documents en se basant sur leur *signature*. Les signatures forment un ordre partiel, similaire à celui des graphes, ce qui permet de sélectionner rapidement les signatures candidates en positionnant la requête dans cette hiérarchie. ELEN a été testé sur un corpus formé de descriptions de méthodes LOOPS et de fonctions du noyau de UNIX.



a) treillis de types b) document «mkdir»

Figure 2.7. Exemple de représentation dans ELEN

e) Extraction & concepts

L'indexation dans RUBRIC (*RUle-Based Retrieval of Information by Computer*) [TAAC87] est toujours réalisée par mots-clés, mais elle est augmentée d'un thésaurus qui définit les concepts du domaine et les relations qu'ils ont entre eux par des structures de *frames*. La figure 2.8, par exemple, montre le concept *meurtre* dans ce système. Les mots en gras et

les lignes continues montrent les concepts et la hiérarchie de concepts. Les mots soulignés et les lignes pointillées montrent les attributs de ces concepts.

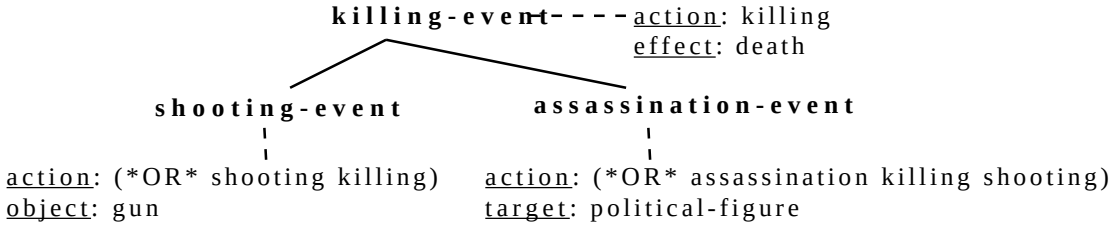


Figure 2.8. Sémantique de «*killing*» dans RUBRIC (tiré de [TAAC87, p.251])

RUBRIC réalise une analyse textuelle pour extraire les *structures évidentielles* du document correspondant aux concepts du thésaurus. Des opérateurs logiques et textuels (proximité, phrase, paragraphe, etc.) sont définis.

RUBRIC a été utilisé entre autres sur un corpus de dépêches d'articles Reuter à propos d'attentats terroristes. Un descendant commercial de RUBRIC, par la firme *Verity Inc.*, est TOPIC, qui lui utilise des articles du *Wall Street Journal*.

L'information est représentée dans FERRET (*Flexible Expert Retrieval of Relevant English Text*) [Mau91] par des *frames*. Cette représentation est similaire donc à ADRENAL, à l'exception près que des *scripts* décrivant les situations les plus courantes sont utilisés lors de l'indexation pour générer les *frames*. L'indexation s'appuie sur une analyse linguistique, FRUMP, qui procède par extraction de concepts dans le texte plutôt que par une analyse syntaxique poussée. FERRET a été testé sur un corpus d'astronomie provenant d'un forum de discussion et de communiqués de l'observatoire McDonald de l'Université du Texas.

LaSSIE [DBS90] utilise KANDOR, un formalisme de représentation de connaissances dérivé de KL-ONE, pour la représentation des index. L'interface permet de traduire automatiquement les requêtes en langue naturelle dans ce formalisme. Une base de connaissances est définie manuellement, et utilisée pour l'indexation. LaSSIE a été testé sur un corpus de génie logiciel consistant en une librairie de composants logiciels réutilisables: le système Definity 75/85 de AT&T. Certaines relations sont extraites automatiquement des codes sources: par exemple les commandes `#defines`, `#includes`, etc.

Enfin, le système HAVANE [BCR85] appartient aussi à cette catégorie. Les *grammaires linguistiques* utilisées dans ce système sont comparables aux structures évidentielles de RUBRIC, et l'information dérivée, des arborescences sémantiques, est de niveau conceptuel.

f) Analyse sémantique & concepts

SCISOR (*System for Conceptual Information Summarization, Organization and Retrieval*) [JR90] utilise KODIAK pour la représentation des documents. Le corpus utilisé dans SCISOR est composé de nouvelles à propos de fusions de corporations (DowJones). L'indexation dans SCISOR est un processus complexe qui implique plusieurs analyses. Une extraction lexicographique est d'abord effectuée. Le texte est segmenté pour empêcher l'analyse de certaines portions jugées trop complexes ou hors-propos. Un filtre est utilisé pour déterminer le thème du document (fusion, acquisition d'une compagnie, etc.). Le processus d'analyse linguistique pour l'indexation est très poussé. Un analyseur ascendant (*bottom-up full parser*) et un analyseur descendant partiel (*top-down skimming partial parser*) sont utilisés. Une caractéristique importante de ce système est qu'il peut résumer les histoires.

L'indexation dans RIME (*Recherche d'Informations MEdicales*) [Ber88, BMB95] a déjà été décrite sommairement dans l'introduction. La représentation se fait par des arborescences sémantiques. L'indexation est réalisée par une analyse syntaxico-sémantique des documents. Le corpus, enfin, consiste en des comptes-rendus médicaux portant sur des images radiologiques.

MENELAS [BZ92], tout comme ELEN, est basé sur les graphes conceptuels. L'indexation est cependant réalisée ici de façon automatique, par le biais d'une grammaire contextuelle et d'une analyse sémantique/pragmatique. Le corpus est constitué de comptes-rendus médicaux portant sur des maladies coronaires. Il s'agit d'une base de données vaste et multilingue (français, anglais, néerlandais). MENELAS utilise aussi une base de connaissances divisée en plusieurs composantes: un catalogue positif, négatif, une liste des référents uniques, ainsi que des schémas ou *scripts* décrivant les situations.

2.4.2 Évaluation des systèmes

Il est difficile d'évaluer les systèmes présentés à la section précédente de façon absolue, puisque chacun présente des avantages et des inconvénients selon le type d'application envisagé. Nous en discutons ici brièvement, en reprenant une à une les catégories selon les deux axes considérés, à savoir l'*indexation* et la *représentation de l'information.*

a) Axe d'indexation

L'indexation *manuelle* a surtout été introduite dans notre grille par souci de complétude, afin de permettre la catégorisation de systèmes présentants d'autres aspects intéressants. Cette approche n'est envisageable que sur des «cas d'école», c'est-à-dire sur un ensemble restreint de documents, et où la cohérence de l'indexeur peut être garantie. D'ailleurs les trois systèmes présentés dans cette catégorie, de l'aveu même de leurs auteurs, n'ont jamais dépassé le stade de prototype.

L'*extraction* est intéressante parce qu'elle requiert peu de ressources pour être mise en œuvre, et peut être réalisée de façon très efficace. Cependant une condition à son utilisation est que le langage employé dans les documents doit contenir suffisamment d'idiomes et d'expressions typiques, afin de permettre l'identification de concepts avec un taux d'erreur acceptable. Ainsi, les systèmes présentés dans cette catégorie répondent tous à ce critère: TOPIC étant appliqué au langage des affaires, LaSSIE a des programmes, etc.

La *lemmatisation* est une solution typique de la recherche d'informations, dont la motivation première est de pallier à des approches linguistiques jugées inapplicables sur de larges volumes de données. Des algorithmes simples, qui ne sont pas dédiés à un domaine d'application particulier, existent pour diverses langues. Son principal inconvénient est l'ajout de bruit; en effet, si son but est normaliser les groupes de mots de même famille, parfois cette normalisation n'est pas souhaitable. Par exemple, le lemme «*inform*» pourrait s'appliquer à la fois à «*informer*» et à «*informatique*». Notons toutefois que l'avènement de dictionnaires électroniques et d'analyseurs morpho-lexicaux performants permet désormais une lemmatisation basée sur la racine des mots, et donc plus fiable.

L'analyse *syntaxique* présente l'avantage, pour un coût relativement modique, de réaliser une analyse plus poussée que la lemmatisation et de permettre le regroupement de structures au sein d'une phrase. Il n'est pas exclu qu'une analyse syntaxique utilise également la lemmatisation, comme c'est le cas pour les systèmes CLARIT et FASIT, bien que les systèmes qui utilisent une représentation plus poussée, lui préféreront généralement une analyse morpho-lexicale, comme pour COP et IOTA. L'inconvénient de l'analyse syntaxique est que, sans l'apport d'un niveau sémantique, elle comporte souvent des ambiguïtés.

L'analyse *sémantique* est celle qui est le mieux à même de répondre au besoin de l'utilisateur, puisqu'elle cherche à comprendre les documents. Là encore, l'emploi d'une analyse de niveau inférieur n'est pas exclue. La coopération se fait le plus souvent dans l'axe syntaxe-sémantique (RIME et SCISOR), mais peut aussi être augmentée d'une collaboration morpho-syntaxique (MENELAS). L'inconvénient majeur de cette analyse, cependant, est sa lourdeur, autant dans la réalisation que du point de vue de l'efficacité. De plus elle est toujours dédiée à un domaine particulier: les applications sont difficilement transférables d'un domaine à l'autre.

b) Axe de représentation de l'information

Le grand avantage des *mots-clés* est leur simplicité; ce qui permet des algorithmes de recherche très performants. Ils peuvent facilement être appliqués à tout type de document et à n'importe quel domaine d'application. Cependant cette représentation est trop pauvre, et est donc exclue pour les systèmes où la précision est importante. D'une part on fait l'hypothèse qu'il y a bivalence entre les mots-clés et la sémantique, alors qu'en réalité un mot en langue naturelle a souvent plusieurs sens possibles. D'autre part on ne modélise pas les relations entre termes. Par exemple, «*la pomme de terre*» aura la même représentation que «*la pomme et la terre*».

Les représentations par *groupes* et de *surface* possèdent une meilleure précision que les mots-clés, et sont indépendants du domaine d'application, ce qui n'est pas le cas pour les structures sémantiques. Le danger ici est de produire des groupements « *artificiels* », c'est-à-dire qui ne réflètent pas la construction mentale de l'utilisateur, ce qui risque de fausser les critères de recherche. De plus, dans certains cas l'analyse syntaxique n'a pas un grand impact sans la prise en compte de sémantique. On n'a qu'à penser à la préposition « *de* » qui s'emploie avec différents sens.

Les systèmes s'appuyant sur une indexation *conceptuelle* permettent une recherche plus précise. Ils sont de plus indépendants de la langue, et sont donc avantageux pour des approches multilingues. Par contre ils sont toujours dédiés à une application ou un domaine particulier. Des projets comme CYC [Len95], qui visent une modélisation des connaissances indépendante du domaine, n'ont eu jusqu'à présent que des résultats mitigés.

2.4.3 Autres aspects

Nous abordons dans cette section trois aspects qui, bien qu'ils n'entrent pas dans notre classification générale, devraient être pris en compte dans un modèle d'indexation. Il s'agit de la pertinence de l'information, du besoin de l'utilisateur, et de la structure des documents.

a) Mesure de l'information

Un des points importants en recherche d'informations est le fait qu'on cherche à *extraire* l'information d'un large volume de données. Cela implique qu'on doit pouvoir comparer différentes informations entre elles, afin de ne présenter à l'utilisateur que les meilleures réponses. De plus, afin de présenter les informations par ordre d'importance, cette comparaison est le plus souvent *valuée*.

Cette mesure de comparaison est appelée la *pertinence*. La notion de pertinence, au-delà de sa définition intuitive, est très difficile à cerner. Elle n'est généralement formalisée qu'au sein d'un modèle de correspondance particulier; cette formalisation existe donc dans des formalismes (logiques terminologiques, graphes conceptuels, etc.) et avec des interprétations diverses (incertitude, croyance, etc.) [Den94]. Deux exceptions notables sont les modèles de [Yao95] et [HD95], qui s'attaquent à la notion de pertinence de l'utilisateur.

Il serait tentant de faire un parallèle avec les autres mesures de l'information que nous avons vues dans ce chapitre: la quantité d'information et le bruit. S'il y a bien sûr, des similitudes, il ne faut pas oublier que contrairement à ces mesures, la pertinence n'est pas *absolue*: elle n'est définie que dans un contexte particulier. Ce contexte dépend bien sûr de la requête et document, mais aussi de l'utilisateur et du système. En fait, comme il est proposé dans [CB96], on peut distinguer au moins deux types de pertinence: la pertinence utilisateur et la pertinence système.

b) Besoin d'information

La notion d'information en recherche d'informations est souvent assimilée au besoin de l'utilisateur. Trois types de besoin utilisateur sont définis dans [Ing92]:

1° Besoin *vérificatif*. L'utilisateur cherche à vérifier le texte avec les données connues qu'il possède déjà. Il recherche donc une donnée particulière, et sait même souvent comment y accéder. La recherche d'un article sur Internet à partir d'une adresse connue serait un exemple d'un tel besoin. Un autre exemple serait de chercher la date de publication d'un ouvrage dont la référence est connue. Un besoin de type vérificatif est dit *stable*, c'est-à-dire qu'il ne change pas au cours de la recherche.

2° Besoin *thématique connu*. L'utilisateur cherche à clarifier, revoir ou trouver de nouvelles informations dans un sujet et domaine connus. Un besoin de ce type peut être *stable* ou *variable*; il est très possible en effet que le besoin de l'utilisateur se raffine au cours de la recherche. Le besoin peut aussi s'exprimer de façon incomplète, c'est-à-dire que l'utilisateur n'énonce pas nécessairement tout ce qu'il sait dans sa requête mais seulement un sous-ensemble. C'est ce qu'on appelle dans la littérature le *label effect*.

3° Besoin *thématique inconnu*. Cette fois, l'utilisateur cherche de nouveaux concepts ou de nouvelles relations hors des sujets ou domaines qui lui sont familiers. Le besoin est intrinsèquement *variable* et toujours exprimé de façon incomplète.

c) Structure de documents

Un dernier facteur qui doit être considéré dans un modèle d'indexation est la structure des documents. Il y a un intérêt accru dernièrement pour cet aspect en recherche d'informations. Pour plusieurs applications travaillant sur le texte intégral, c'est-à-dire sur des textes entiers et non pas seulement sur des résumés ou des extraits, il semble nécessaire de pouvoir au moins reconnaître la structure du texte, de façon à pouvoir distinguer par exemple, un titre du corps du texte. Pour les systèmes travaillant sur des corpus spécialisés, il est souvent aussi nécessaire d'inclure la structure dans la représentation du document, puisqu'elle est parfois un aide indispensable à la compréhension du document.

Un des premiers systèmes intégrant la structure dans le processus d'indexation fut IOTA [Ker84]. La structure dans IOTA est utilisée principalement pour donner un meilleur contrôle sur la précision et le rappel. Ceci est rendu possible par le biais d'une unité d'indexation *dynamique*, c'est-à-dire en contrôlant les bornes inférieures et supérieures de la taille d'un terme d'indexation. Ces niveaux sont choisis a priori pour un type de document donné; ainsi pour l'expérimentation avec le corpus technique du CNET, qui a servi à valider l'approche, le niveau maximal était le chapitre, et le niveau minimal, le paragraphe.

Des règles pour calculer la pertinence d'un terme d'une section inférieure du texte à une section supérieure sont définies. Cette stratégie est appelée la *remontée des termes*.

En particulier, les termes d'indexation d'un titre ou de certaines sections sont propagées à leur section mères.

Un autre exemple d'utilisation de la structure dans un système de recherche d'informations est trouvé dans [TSM91]. Dans cette approche, la structure du corpus est définie par une grammaire, et le processus d'indexation est également décrit par une grammaire utilisant celle du corpus. Les documents peuvent être accédés par un hypergraphe, où les nœuds représentent les parties de documents, et les liens représentent les références d'une partie à une autre. Cette approche a été conçue et testée sur le corpus de *Canadian Patent Reporter*, qui consiste en des rapports en texte intégral de procès ayant trait aux brevets d'invention et aux marques de commerce.

Des travaux plus récents ([Cal94],[Wil94]) concernent essentiellement le classement des documents structurés à partir des passages retrouvés (*passage retrieval*).

Une autre approche est l'utilisation d'hypertextes[15] [Dun91], [Sav94] ou [Khe95]. Les hypertextes considèrent typiquement deux types de relations dans l'expression des documents: les liens de composition – c'est-à-dire comment un document est formé à partir de sa structure – et les liens relationels – ou les liens sémantiques entre les parties d'un document. Ces relations peuvent encore être raffinées, comme dans [KC95], où des relations comme «même-sujet», «illustré-par», etc., sont considérées.

Un exemple de combinaison hypertexte/recherche d'informations est donné dans [KC95]. Cette approche nous intéresse surtout quant à son utilisation de la structure logique des documents. Les documents sont représentés par une arborescence où les index des sections supérieures sont calculés par combinaison des index de leurs sous-sections. La recherche s'effectue par application de la fonction d'implication directe $D \rightarrow Q$ sur les sections supérieures du document (opération *fetch*). Quand une de ces sections est retrouvée, on peut raffiner la réponse en appliquant la même fonction à ses sous-sections (opération *browse*).

Malheureusement, ces travaux, bien qu'ils reconnaissent l'importance de la structure, sont habituellement dépendants d'une application. Par exemple, le modèle décrit dans [TSM91] est très dépendant de la structure spécifique aux documents légaux.

2.5 L'information dans un modèle d'indexation

Nous avons présenté dans les sections précédentes les différents sens qu'on a donné à l'information, d'une part dans les théories de l'information et de la linguistique, et d'autre part en recherche d'informations. Si les premiers travaux présentent un intérêt certain du point de vue théorique, puisqu'ils permettent de mieux comprendre certains aspects psycholinguistiques, ils demeurent assez éloignés du sens pratique de l'information en recherche d'informations, et sont malheureusement souvent trop abstraits pour faire l'objet d'une implémentation informatique. D'un autre côté, les travaux de recherche d'informations,

15. Une bonne introduction aux hypertextes peut être trouvée dans [Con87].

en faisant des données pragmatiques le cœur de leurs préoccupations, négligent l'aspect théorique, et de ce fait ont une conception de l'information qui ne correspond pas à l'idée intuitive qu'en a l'utilisateur.

Il nous semble donc primordial de faire le pont entre ces deux mondes, tâche que nous comptons réaliser en combinant les différents aspects dans un même cadre. Nous ne proposons pas une définition ou une interprétation unique du concept d'information, mais plutôt une classification des différents types d'informations pouvant être utiles dans un modèle d'indexation.

Après avoir justifié à la section suivante cette approche, nous introduisons les différents types d'information, qui seront décrits en détails au chapitre 3. Nous concluons par une analyse d'une notion centrale à la recherche d'informations, la *pertinence*, par rapport aux types d'informations définis.

2.5.1 Types de requête

L'analyse des systèmes de recherche d'informations que nous avons effectuée à la section précédente montre que la tendance actuelle semble être vers une représentation plus complexe de la sémantique des documents, de leur structure, et éventuellement de l'emploi de connaissances.

Si une représentation sémantique plus poussée est appropriée dans bien des cas, il faut voir que, comme tout type de représentation, elle ne peut répondre qu'à certains types de requête. Considérons par exemple les trois requêtes suivantes:

1° Chercher un document au titre comme «*Le dernier des Mohicans*»;

2° Chercher un document à propos des *Mohicans*;

3° Chercher un document où les sœurs Monro sont enlevées par des Mohicans.

Ces requêtes font toutes trois référence au livre «*Le dernier des Mohicans*» de Fenimore Cooper. Chacune fait aussi référence à un élément d'information distinct. La requête n° 1 réfère au titre du volume, la seconde au thème de ce texte, et la troisième à son sens. Une représentation explicitant la sémantique du document, puisqu'elle a perdu tout lien avec la structure linguistique du texte, ne saurait répondre à la requête n° 1. Elle ne peut répondre que de façon imparfaite à la requête n° 2, puisque certains indices pour l'extraction du thème, qui sont donnés par les structures linguistiques et de discours, sont absents de cette représentation.

Les requêtes suivantes font appel à d'autres éléments d'informations du même livre:

4° Chercher le célèbre livre écrit par Fenimore Cooper;

5° Chercher le chapitre où le massacre de William Henry est décrit;

6° Chercher la définition de «*Narragansett*» [16].

Le langage d'indexation doit donc non seulement représenter l'information qui est recherchée par l'utilisateur – il s'agit pour les requêtes ci-dessus du livre lui-même ou de l'un de ses passages – mais il doit aussi – pour ne pas dire surtout! – représenter l'information complémentaire qui peut aider à retrouver cette information. Un besoin de type *thématique connu*, comme nous l'avons introduit à la section 2.4.3b), fait habituellement appel à cette information complémentaire. Par exemple, dans la requête n° 2, l'utilisateur n'est pas intéressé par le thème du document, qu'il connaît déjà, mais bien par le document lui-même.

L'information complémentaire peut figurer explicitement dans la requête, comme c'est le cas pour les requêtes ci-dessus, ou être utilisée de façon transparente à l'utilisateur, comme c'est souvent le cas pour l'utilisation de la *structure* ou des *connaissances* (voir section suivante).

Ces quelques exemples devraient convaincre le lecteur de la nécessité de prendre en compte divers *types* d'information dans un modèle d'indexation, puisque les requêtes peuvent faire appel, directement ou indirectement, à différents types d'informations. Nous nous employons dans la prochaine section à présenter ces types.

2.5.2 Types d'information

La table 2.3 résume les différents types d'information qui peuvent être considérés dans un modèle d'indexation pour le texte. Les types d'information sont classés selon deux axes: un premier axe que nous qualifions de *linguistique*, puisqu'il reprend les aspects signifié, signifiant et pragmatique, et un second axe qualifié de *représentationnel*, qui s'adresse plus particulièrement aux problèmes de représentation propre à la recherche d'informations.

La première catégorie sous l'axe *représentationnel* est celle du *contenu* en lui-même, qui correspond au discours du document. Le *méta-contenu*, c'est l'«information sur l'information», ou en d'autres termes, tout ce qui est extérieur au contenu mais qui s'y rattache de façon directe ou indirecte. Enfin, la *structure du contenu* s'intéresse à la façon dont est organisée l'information, c'est-à-dire à la segmentation en parties du document et aux différents liens de structure entre les éléments d'information.

De façon plus intuitive, ces trois catégories peuvent être vues comme l'information répondant respectivement aux trois questions suivantes:

- «*De quelle information est-il question dans le texte?*»
- «*Que dit-on (ou que sait-on d'autre) sur cette information?*»

16. Une race de cheval du Nouveau Monde.

– « *Comment cette information est-elle organisée (par rapport aux autres informations) dans le texte?* »

L'axe *linguistique* a déjà été discuté dans ce chapitre. Nous établissons clairement la distinction entre signifiant et signifié, distinction qui n'apparaît pas toujours clairement en recherche d'informations. Cette frontière est nécessaire afin de pouvoir réutiliser les travaux de linguistique et de théories du discours. Par contre, le signifié et la pragmatique sont regroupés dans notre classification: ces deux niveaux s'adressent tous deux à la sémantique, la différence ne se situant pas tant dans l'information elle-même, mais plutôt dans les procédés utilisés pour la dériver.

L'information donnée explicitement par les documents ou par leurs représentations électroniques est généralement de l'ordre du signifiant. C'est malheureusement l'information qui nous intéresse le moins, puisque les utilisateurs sont souvent plus intéressés par la signification que par les symboles eux-mêmes. C'est donc là une des tâches essentielles du modèle d'indexation: dériver le signifié à partir du signifiant.

Les sections suivantes décrivent brièvement les types présentés à la table 2.3, et par le fait même donnent les grandes lignes du langage d'indexation qui sera défini dans les chapitres à venir. Il va sans dire qu'une application donnée ne va pas nécessairement représenter tous les types d'information; tout dépend des requêtes qui pourront être formulées dans le système et des documents qui forment le corpus.

Table 2.3. Types d'information dans un modèle d'indexation

	signifiant	*signifié/pragmatique*
contenu	contenu textuel	contenu sémantique
méta-contenu	nature du contenu attributs de contenu	thèmes connaissances
structure du contenu	structure linguistique structure logique	structure de discours

a) Contenu textuel

On entend par *contenu textuel* tout texte qui apparaitrait dans une version imprimée du document. En d'autres termes, c'est tout ce qui est destiné à être lu par le lecteur final du document, par opposition aux agents intermédiaires comme le processus d'indexation. Il s'agit donc du corps du discours, bien entendu, mais aussi d'autres informations comme les titres de sections, les figures et les tables, etc.

Le contenu textuel s'apparente au *message* de Shannon, tel que présenté à la section 2.2.1, l'émetteur étant ici l'auteur du document, et le récepteur, le lecteur.

Dans nos exemples de requêtes ci-dessus, la requête n° 1 fait appel au contenu textuel, puisqu'elle ne s'adresse pas à la sémantique, mais uniquement à la forme superficielle du texte.[17]

b) Contenu sémantique

Si la distinction entre *contenu textuel* et *contenu sémantique* est assez évidente, puisqu'il s'agit simplement de l'opposition signifiant/signifié, il convient cependant de différencier le contenu sémantique du thème. Bien que tous deux appartiennent au signifié, le contenu sémantique est généralement beaucoup plus complexe, mettant en scène différents concepts et relations entre ces concepts. Il reflète le sens qui est véhiculé par le texte, et correspond donc plus ou moins fidèlement au schéma mental que se ferait un lecteur après lecture de ce texte. Le thème, quant à lui, répond à la question « *de quoi parle-t-on dans ce texte?* », et est donc à voir comme une information supplémentaire à propos du texte.

Le contenu sémantique ne correspond donc pas aux *attributs internes* tels qu'ils sont utilisés traditionnellement en recherche d'informations, puisque ces derniers incluent généralement la notion de thème.

Le contenu textuel s'apparente au *contenu* de l'information sémantique, tel que présenté à la section 2.2.2, bien que sa représentation ne soit pas limitée à des propositions logiques.

La requête n° 3 fait appel au contenu sémantique, en effet, pour répondre à cette requête, il faut avoir compris le sens du document.

c) Nature du contenu

On entend par *nature du contenu* toute *caractérisation* du contenu textuel qui en spécifie le rôle fonctionnel, logique, ou même sémantique. Dans un formalisme comme SGML, cela correspond au *marquage* du texte.

Ainsi, le terme «*Narragansett*» rencontré dans un document pourra être identifié comme un *terme technique* (marquage *sémantique*). À l'endroit où il apparaît sur la page couverture, «*Le dernier des Mohicans*» peut être identifié comme le *titre* du document (marquage *logique*). Enfin, «*les sœurs Monro*» peut être identifié comme *groupe nominal* (marquage *fonctionnel*). En supposant ce marquage, les requêtes n° 1, n° 3 et n° 6 font appel à la nature du contenu.

17. Cette requête fait également appel à la *nature du contenu*, comme nous le verrons ci-dessous.

d) Attributs de contenu

Les attributs de contenu sont des informations complémentaires qui portent sur le contenu textuel. On regroupe ici aussi bien les attributs s'adressant aux symboles eux-mêmes, comme par exemple la langue ou l'alphabet, que ceux portant sur le texte, comme par exemple l'auteur d'une citation, le développement d'une abbréviation, etc.

Les attributs bibliographiques, traditionnellement appelés *attributs externes* dans la littérature, sont également de ce type, à cette différence près qu'ils se rapportent à tout un document au lieu d'un passage du texte. D'autre part, plusieurs de ces attributs peuvent aussi correspondre à des éléments du contenu textuel. Nous avons vu par exemple que «*Le dernier des Mohicans*», s'il apparaît sur la page couverture, peut déjà être identifié comme *titre*. Le fait de disposer d'un attribut bibliographique distinct de type *titre* permet entre autres de pouvoir différencier au besoin le titre qui apparaît dans la version imprimée de celui utilisé pour la recherche ou l'indexation. Par exemple, on pourrait donner ici comme titre *bibliographique*: «*Le dernier des Mohicans: version électronique*».

La requête n° 4 fait appel à un attribut de contenu: l'auteur du document.

e) Thème

Nous considérons le thème comme notion centrale dans notre modèle. Nous avons déjà vu trois interprétations possibles du thème:

- ce dont on parle dans une phrase (analyse statutaire à la section 2.3.1);
- ce qui est supposé connu dans une communication (progression thématique à la section 2.3.1);
- ce que l'énonciateur d'un discours cherche à véhiculer (intentions à la section 2.3.2).

En réalité, ces trois notions expriment plus ou moins la même chose [Car83] mais à des niveaux différents. Le thème de l'analyse statutaire est au niveau intra-phrase, le thème de la progression thématique est au niveau inter-phrase, et le thème-intention est au niveau du discours. Notre notion de thème inclut donc ces trois visions.

La requête n° 2 fait référence au thème.

f) Connaissances

Ce type d'information s'applique au signifié. Nous avons déjà vu à la section 2.4.1 plusieurs formes de connaissances: taxonomie de concepts, règles d'inférence, thésaurus, etc.

La requête n° 4, par exemple, présuppose qu'on sache déduire quel est le plus célèbre des livres écrits par Fenimore Cooper, ce qui peut être réalisé par une connaissance externe portant sur cet auteur.

g) Structures linguistiques et logiques

Nous avons discuté à la section 2.4.3c) de l'importance de la structure, qui se traduit dans notre modèle par la distinction de trois types de structures: la structure linguistique, la structure logique, et la structure de discours.

La structure linguistique indique la structure superficielle ou syntaxique des symboles formant le texte. La structure logique, elle, indique la composition du texte en chapitre, sections, etc., ainsi que les références dans le texte à des passages ou à d'autres documents.

La requête n° 5 fait directement référence à un information de structure logique: la composition du document en chapitres.

h) Structure de discours

La structure de discours correspond à l'organisation des idées dans le texte, de façon similaire à la structure intentionnelle montrée à la section 2.3.2. Elle est généralement reflétée par la structure hiérarchique logique – puisqu'autrement les textes seraient incompréhensibles! – mais y apporte plus de détails, comme par exemple une segmentation inter-paragraphe.

Peu de requêtes font directement appel à ce type d'information, par contre, elles sont très utiles pour dériver les thèmes, comme nous le verrons au chapitre 4.

2.5.3 Facteurs pour la pertinence

En recherche d'informations, la *pertinence* est une mesure qui indique comment un document satisfait à une requête. C'est donc avant tout une mesure qui intervient lors de la recherche, c'est-à-dire quand la correspondance entre documents et requête est effectuée. Toutefois, de nombreux facteurs intervenant dans l'évaluation de la pertinence sont indépendants de la requête, et peuvent être calculés *a priori*. On a tout avantage à bien isoler ces facteurs, et à les évaluer quand c'est possible lors de l'indexation des documents.

- L'**incertitude** représente le degré de confiance que l'on accorde à un élément d'information. Cette notion ne doit pas être confondue avec d'autres facteurs de pertinence, ou pire encore, avec la pertinence elle-même: on exprime par l'incertitude le doute quant à la *nature* même d'un élément d'information. Par exemple, on peut douter du texte lui-même à cause du processus de saisie ou de reconnaissance automatique

de caractères, ou de la représentation sémantique à cause des phénomènes de paraphrasage ou de polysémie, etc. En fait, l'incertitude est sans doute le facteur de pertinence le plus universel, puisqu'il peut s'appliquer à tout élément d'information. Son interprétation peut varier, cependant, d'un type d'information à un autre.

- L'**à-propos** (*aboutness*) mesure à quel point un thème est représentatif d'un passage du document. Contrairement à l'incertitude, qui est une mesure absolue, i.e. qui prend son sens sans avoir à se rattacher à d'autres éléments, cette mesure est toujours relative à une structure textuelle. Par exemple, si cette section est bien à-propos de «*pertinence*», la section 2.5 l'est à un degré moindre puisqu'il ne s'agit plus du thème principal.

- La **distance sémantique** est une mesure qui évalue à quel point deux *signifiés* sont reliés. On peut ainsi comparer les représentations sémantiques, les thèmes, ou les connaissances. Par exemple, «*cabane*» et «*maison*» ont un certain lien entre elles, qui sera représenté par la distance sémantique.

- La **saillance** est une mesure de l'importance d'un signifiant – une structure linguistique ou logique – par rapport à un passage du document, de par sa position ou son rôle dans ce passage. Cela correspond intuitivement aux structures qui sont considérées comme les plus porteuses d'information dans un document quel que soit leur contenu. Imaginons par exemple qu'on vous demande de résumer un livre, mais que vous ne disposiez de peu de temps. Il est probable que les éléments d'information auquel vous attacherez le plus d'importance seront: le titre, le résumé, bien sûr s'il existe, les titres de chapitres/sections, l'introduction et la conclusion, les premières lignes des paragraphes, etc. Ces éléments sont choisis parce qu'ils sont plus *saillants*.

Différentes mesures peuvent s'appliquer à un même élément d'information. Considérons par exemple le thème «*pertinence*» pour cette section. Peut-être, dû à une mauvaise interprétation du texte, ne fallait-il pas le considérer comme thème? (facteur d'incertitude) Peut-être n'en constitue-t-il qu'un thème secondaire? (facteur d'à-propos) Peut-être, le terme «*relevance*» serait lui aussi approprié? (facteur de distance sémantique) Enfin, le choisissons-nous comme thème sans même avoir lu la section, uniquement parce qu'il apparaît dans le titre de la section? (facteur de saillance)

Il est clair que cette liste n'est pas exhaustive, mais elle forme une base de travail qui sera utile pour notre modèle d'indexation. D'autres facteurs qui auraient pu être considérés sont: l'importance d'un attribut externe, l'importance d'une connaissance dans un contexte, etc.

2.6 Conclusion

Notre problématique dans ce chapitre était de clarifier la notion d'information textuelle pour la recherche d'informations, tâche essentielle avant de pouvoir définir un modèle d'indexation. Il ressort de notre analyse que l'information textuelle peut être considérée de différents points de vue, tant par des disciplines diverses, qu'au sein même de la recherche d'informations. Nous avons donc tenté dans la proposition d'un modèle d'information, de combiner plusieurs de ces vues, en définissant des types d'information.

Six grands types d'information sont définis, qui peuvent chacun être vu selon deux aspects: l'aspect linguistique ou l'aspect représentationnel. Un autre critère, à savoir la *présentation* (l'aspect physique du texte sur le papier ou à l'écran, ou le *layout*) est ignoré. Bien qu'elle ait sans aucun doute un impact sur l'appréhension des textes, et donc sur la recherche d'informations, la présentation constitue un problème à part, que nous ne traitons pas ici.

Les types d'information définis permettent en outre de clarifier le rôle et la position de certaines notions qui deviennent de plus en plus floues à mesure que la représentation des documents devient plus complexe, à savoir: la signification, le thème, la structure, et les connaissances. Enfin, la notion de pertinence par rapport à cette définition d'information a été examinée.

Chapitre 3

Comment représenter l'information

La connaissance conduit à l'unité
comme l'ignorance mène à la diversité.

Gadâdhar Chatterji, dit RÂMAKRISHNA
(*Entretiens*)

NOUS PRÉCISONS dans ce chapitre quelles informations sont utiles pour la recherche d'informations; comment elles se manifestent dans les textes et comment les représenter de façon cohérente. Pour ce faire, nous définissons le langage de représentation $\mathcal{L}$, qui permet de représenter à la fois le contenu d'un document, ses index, ainsi que toute information permettant de dériver les index.

Nous présentons d'abord les grandes lignes du langage et abordons quelques problèmes de représentation. Les éléments du langage sont organisés par types d'information, tels que définis au chapitre précédent: nous décrivons successivement les éléments de *contenu*, de *méta-contenu*, et de *structure de contenu*. Enfin, nous concluons en discutant de l'intégration de ces types d'information, à l'aide d'un exemple complet de document.

3.1 Aperçu de $\mathcal{L}$

Le langage de représentation $\mathcal{L}$ doit pouvoir représenter à la fois des informations de l'ordre du signifiant et du signifié. Ainsi, nous nous inspirons pour notre tâche de formalismes issus du monde de l'édition électronique ou de l'échange de documents, comme par exemple: LaTeX, RTF, ODA, SGML, etc., qui s'adressent au signifiant, ou à des langages de représentation de connaissances comme les graphes conceptuels, logiques terminologiques, etc., qui eux s'adressent au signifié.

Le langage que nous proposons n'a pas la prétention de remplacer les standards ou langages existants: il nous permet d'exprimer les différentes informations sous une même

forme, et ce de la façon la plus intuitive possible. Les documents source existent sous des formats très variés: le langage $\mathcal{L}$ permet de garder le modèle d'indexation indépendant de ces formats.

L'information dans $\mathcal{L}$ est organisée selon la typologie déjà proposée au chapitre précédent (cf. table 2.3). La table 3.1 montre les principaux prédicats de $\mathcal{L}$ selon cette classification.[1] La base de la représentation d'un document est formée par son *contenu-signifiant* (désigné par σ dans la table), c'est-à-dire les passages du document auquels sont associés le texte (τ). Un *contenu-sémantique* (ς), qui donne le sens du passage, peut être associé au contenu-signifiant. Ces deux types de contenu peuvent être raffinés en leur ajoutant un type *méta-signifiant* ou *méta-sémantique*. Enfin, d'autres informations de l'ordre du méta-contenu et de la structure sur le contenu sont ajoutées au contenu.

Table 3.1. Classification des prédicats dans $\mathcal{L}$

	signifiant	signifié/pragmatique
contenu	*contenu-signifiant*(σ) *texte*(σ, τ)	*contenu-sémantique*(ς) *sém*(σ, ς)
méta-contenu	*méta-signifiant*(σ) *auteur*(σ, α), *titre-doc*(σ, α), *date*(σ, α) *publié-par*(σ, α), *publié-à*(σ, α), *édité-par*(σ, α) *sélection*(σ, α), *id*(σ, α), *auteur-citation*(σ, α) *expan*(σ_1, σ_2), *abbr*(σ_2, σ_1), *langue*(σ, α) *alphabet*(σ, α), *externe*(σ, τ), *niveau*(σ, α) *lignes*(σ, α), *colonnes*(σ, α), *nom-agent*(α, τ) *valeur-temps*(α, τ)	*méta-sémantique*(ς) *thème*(σ, ς) *équivalent*$(\varsigma_1, \varsigma_2)$ *implique*$(\varsigma_1, \varsigma_2)$
structure du contenu	*part*(σ_1, σ_2), *part-trans*(σ_1, σ_2) *séq*(σ_1, σ_2), *séq-trans*(σ_1, σ_2) *adjacent*(σ_1, σ_2), *réf*(σ_1, σ_2) *désigne*(σ_1, σ_2), *dépendant*(σ_1, σ_2) *contient-texte*(σ, τ)	*domine*(σ_1, σ_2) *domine-trans*(σ_1, σ_2) *intention*(σ_1, σ_2)

Les prédicats de la table 3.1, qui sont l'objet principal de ce chapitre, sont détaillés dans les sections 3.2 à 3.4. Nous illustrons l'emploi de ces prédicats et définissons des

1. Le prédicat *alternative*, qui peut apparaître à tous les niveaux, est omis de cette table. Les prédicats suivants, qui servent à la définition de $\mathcal{L}$ mais pas directement à l'expression de l'information, sont également omis de la table: *chaîne*, *nombre*, *transitive*, *réflexive*, *irréflexive*, *symétrique*, *antisymétrique*, et *injective*.

propriétés sur leurs contraintes d'utilisation. Avant d'aborder cette description, toutefois, il nous semble essentiel de définir formellement la syntaxe et la sémantique de $\mathcal{L}$ (sections 3.1.1 et 3.1.3). Les principaux problèmes de représentation sont aussi discutés (section 3.1.2).

3.1.1 Syntaxe de $\mathcal{L}$

La syntaxe de $\mathcal{L}$ est essentiellement une extension à la syntaxe de la logique des prédicats, où les prédicats peuvent être accompagnés par une mesure de *certitude* ou de *poids*. En plus des connecteurs habituels, conjonction, négation, modus ponens, et des quantificateurs universels et existentiels, un opérateur d'égalité est également défini.

Le langage permet d'exprimer des *faits*, qui, lorsque combinés, forment la définition des documents, ainsi que des *règles*, qui expriment des contraintes plus générales sur l'usage des prédicats.

a) Faits

Un `Fait` correspond à un `Prédicat` qui exprime une information relative à un document ou à une partie de document: il peut s'agir du texte associé à un passage, de la composition structurelle d'un chapitre, d'attributs bibliographiques, etc. La représentation d'un document est donnée par l'ensemble des faits qui s'y rattachent. Par exemple, le document $d1$ ci-dessous est défini par deux faits, qui donnent son type, un *rapport-technique*, et son auteur.

rapport-technique(*d1*).
auteur(*d1*, «*François Paradis*»).

La syntaxe d'un fait est donnée par la BNF ci-dessous:[2]

2. Les conventions suivantes sont adoptées:

- la barre verticale (|) représente des *alternatives*, les crochets ([ ]) un élément optionnel, et les accolades ({}*) un élément répété zéro ou plusieurs fois.

- les éléments entre guillemets (‘ ’) ou en *italique* correspondent aux éléments terminaux du langage.

```
Fait ::= Prédicat '.'
Prédicat ::= Prédicat-C | Prédicat-P
Prédicat-C ::= Prédicat-C1 | Prédicat-C2
Prédicat-C1 ::= Nom-Prédicat-C1 '(' Arg ')' [ '[' Certitude ']' ]
Prédicat-C2 ::= Nom-Prédicat-C2 '(' Arg ',' Arg ')' [ '[' Certitude ']' ]
Prédicat-P ::= Nom-Prédicat-P '(' Arg ',' Arg ')' [ '{' Poids '}' ]
```

Deux types de prédicats sont définis dans $\mathcal{L}$: les `Prédicat-C` sont les informations auxquelles s'attachent une *certitude*, alors que les `Prédicat-P` sont celles auxquelles s'attachent un *poids*. Les prédicats de certitude sont des relations pour lesquelles on peut exprimer une «*confiance*» sur la valeur de vérité, tandis que pour les prédicats de poids, on exprime plutôt dans quelle mesure la relation entre deux éléments est vérifiée. Par exemple, imaginons deux prédicats *homme* et *grand* qui indiquent respectivement si un individu est de sexe masculin ou s'il est grand. Le prédicat *homme* est forcément vérifié ou non pour un individu donné, mais il peut l'être avec une certaine marge d'erreur; on peut ne connaître que les initiales de la personne, ou en avoir une photo floue. Il s'agit donc d'un prédicat de certitude. Par contre, le prédicat *grand* est très relatif; un individu pourra paraître plus ou moins grand dépendamment de la référence de «*grandeur*» ou de l'observateur. Il s'agit d'un prédicat de poids.

Les `Prédicat-C` sont unaires (`Prédicat-C1`) ou binaires (`Prédicat-C2`). Les `Prédicat-C1` permettent essentiellement de *typer* l'information; leur liste partielle est donnée par la règle `Nom-Type` ci-dessous. Le langage $\mathcal{L}$ comporte plus de 100 types, qu'il serait trop long d'énumérer ici, mais qui sont détaillés dans les sections à venir. Les autres `Prédicat-C1` sont des propriétés de constantes (*chaîne*, *nombre*) ou de relations (*transitive*, *réflexive*, etc.).

```
Nom-Prédicat-C1 ::= variable | Nom-Type | 'chaîne' | 'nombre' | 'transitive' |
                    'réflexive' | 'irréflexive' | 'symétrique' |
                    'antisymétrique' | 'injective'
Nom-Type ::= 'contenu-signifiant' | 'contenu-sémantique' |
             'méta-signifiant' | 'méta-sémantique' | etc.
```

Les `Prédicat-C2` expriment une relation quelconque entre deux éléments d'information, qui peut lier un élément de contenu textuel à un `Attribut`, exprimer la structure, etc.

`Nom-Prédicat-C2` ::= *variable* | `Attribut` | '*adjacent*' | '*contient-texte*' | '*désigne*' | '*domine*' | '*domine-trans*' | '*intention*' | '*non-séq*' | '*part*' | '*part-trans*' | '*réf*' | '*sém*' | '*séq*' | '*séq-trans*' | '*texte*'

`Attribut` ::= '*abbr*' | '*alphabet*' | '*auteur*' | '*sélection*' | '*colonnes*' | '*date*' | '*édité-par*' | '*expan*' | '*externe*' | '*id*' | '*langue*' | '*niveau*' | '*nom-agent*' | '*publié-par*' | '*publié-à*' | '*lignes*' | '*valeur-temps*' | '*titre-doc*' | '*auteur-citation*'

Les `Prédicat-P` sont: *alternative*, *thème*, *équivalent*, *implique* et *dépendant*.

`Nom-Prédicat-P` ::= *variable* | '*alternative*' | '*thème*' | '*équivalent*' | '*implique*' | '*dépendant*'

Les arguments (`Arg`) d'un prédicat sont donnés par une constante ou une variable. Les `Constante`s sont des références à des éléments d'information (*constante-id*), des certitudes (*constante-certitude*), des poids (*constante-poids*), des noms de prédicats (`Constante-Prédicat`), des nombres (*constante-numérique*) ou des chaînes de caractères (*constante-chaîne*).

`Arg` ::= *variable* | `Constante`
`Constante` ::= *constante-id* | *constante-certitude* | *constante-poids* | `Constante-Prédicat` | *constante-numérique* | '«' *constante-chaîne* '»'
`Constante-Prédicat` ::= `Nom-Prédicat-C1` | `Nom-Prédicat-C2` | `Nom-Prédicat-P`

Les *variable*s, quant à elles, sont des symboles pouvant au besoin être suivis d'un indice, qui remplacent une constante quelconque. En pratique, les variables sont de peu d'utilité pour l'expression des faits, et sont essentiellement utilisées dans l'expression des règles.

Afin de différencier les variables des constantes, nous utilisons les lettres grecques pour les variables. Les conventions suivantes sont utilisées:[3]

- ρ pour le nom d'un prédicat: on peut préciser la nature du prédicat en utilisant ρ_c pour un `Prédicat-C`, ρ_t pour un *type*, ou ρ_p pour un `Prédicat-P`;
- σ pour une référence à une information de type *contenu-textuel*;
- ς pour une référence à une information de type *contenu-sémantique*;
- τ pour du texte (chaîne de caractères);

3. Il ne s'agit que de conventions syntaxiques qui ont pour but de clarifier les énoncés.

- μ pour une certitude;
- δ pour un poids;
- α,β pour d'autres variables.

Les prédicats ci-dessous sont des exemples de faits dans $\mathcal{L}$:

part(s, α).
chapitre(s).
texte$(s,$ «*introduction*»$)[.8]$.

Dans cet exemple, s est une *constante-id*, *introduction* est une *constante-chaîne*, *0.8* est une *constante-certitude*, *part*, *chapitre* et *texte* sont des `Constante-Prédicat`, et enfin, α est une *variable*.

Enfin, les mesures de `Certitude` et de `Poids` s'expriment immédiatement après le prédicat, entre des crochets («*[]*») pour les certitudes et des accolades («*{ }*») pour les poids. Ainsi, dans l'exemple ci-dessus, une certitude de .8 est associée au prédicat *texte*.

`Certitude` ::= *variable* | *constante-certitude*
`Poids` ::= *variable* | *constante-poids*

Les certitudes ne prennent pas nécessairement des valeurs numériques, pour autant qu'elles soient ordonnées, c'est-à-dire comprises à l'intérieur d'un intervalle donné. En d'autres termes, un ordre total est défini sur ces mesures, c'est-à-dire que tout couple de valeurs, est réflexif, anti-symétrique, et transitif. Pour les certitudes, l'intervalle est donné par $]\mu_\perp, \mu_\top]$, où $\mu_\perp$ représente la certitude *minimale*, et $\mu_\top$ la certitude *maximale*. De la même façon, l'intervalle pour les poids est donné par $]\delta_\perp, \delta_\top]$.[4] En principe, ces intervalles sont à définir lors de l'implémentation du modèle. Toutefois dans ce travail, à moins d'indication contraire, nous supposerons l'intervalle $]0, 1]$, commun aux certitudes et aux poids.

Lorsqu'elle est omise, la certitude est supposée égale à $\mu_\top$. De même, un poids qui est omis est supposé égal à $\delta_\top$.[5] En supposant $\mu_\top = \delta_\top = 1$, les deux premiers faits de l'exemple ci-dessus peuvent donc être ré-écrits comme suit:

part$(s, \alpha)[1]$.
chapitre$(s)[1]$.

4. L'intervalle ouvert sur $\mu_\perp$ et $\delta_\perp$ implique que les prédicats de certitude ou de poids nul ne sont pas exprimables dans $\mathcal{L}$, ce qui signifie qu'on ne déclare dans $\mathcal{L}$ que des faits possédant un minimum de plausibilité.

5. Cette propriété est décrite plus formellement à la section 3.1.2; règles 3.7 et 3.9.

Un prédicat d'une certitude donnée implique nécessairement ce même prédicat avec un certitude inférieure. Cela correspond à l'idée intuitive que si un fait est connu avec une certaine certitude, alors il est forcément aussi connu avec une certitude moindre. De toute évidence, seule l'expression du prédicat avec sa certitude maximale a besoin d'être conservée dans la base de faits, puisque toutes les autres peuvent en être déduites.[6] Ainsi, toujours pour le même exemple, on peut déduire (entre autres):

$chapitre(s)[.5]$.
$texte(s, \text{«}introduction\text{»})[.1]$.

b) Règles

Les `Règle`s sont utilisées dans notre langage pour exprimer les contraintes d'utilisation des prédicats. De façon plus générale, elles permettent de définir les documents «*valides*» dans notre modèle.

La BNF ci-dessous donne la syntaxe pour une règle.

```
Règle ::= Formule '.'
Formule ::= Prédicat |
            '¬' Formule |
            Arg '=' Arg | Arg '≠' Arg |
            Certitude '≤' Certitude | Poids '≤' Poids |
            Formule '∧' Formule | Formule '∨' Formule |
            Formule '⊃' Formule | Formule '−' Formule |
            '∃' VarListe Formule | '∀' VarListe Formule
VarListe ::= variable { ',' variable } *
```

Les règles sont basées sur l'implication matérielle et l'équivalence logique, notées respectivement par «$\supset$» et «$-$». Elles peuvent comporter les connecteurs logiques habituels de négation («$\neg$»), de conjonction («$\wedge$») ou de disjonction («$\vee$»).

L'égalité («$=$») et la non-égalité («$\neq$») sont définies entre des `Arg`s, c'est-à-dire des constantes ou des variables. L'inégalité («$\leq$») n'a de sens qu'entre les certitudes ou les poids, puisque seules ces constantes sont ordonnées.

Les variables dans une règle peuvent être quantifiées universellement («$\forall$») ou existentiellement («$\exists$»). Par défaut, les variables libres sont quantifiées universellement. Ainsi, les deux règles ci-dessous sont équivalentes:

$$part(\sigma_1, \sigma_2) \quad \supset \quad (\sigma_1 \neq \sigma_2).$$

6. Voir les règles 3.8 et 3.10.

$$\forall \sigma_1, \sigma_2 \; part(\sigma_1, \sigma_2) \supset (\sigma_1 \neq \sigma_2).$$

Cette quantification implicite des variables permet d'alléger les règles.

Les propriétés habituelles de transitivité, réflexivité, etc. sur les prédicats sont définies comme suit:

$$transitive(\rho) - (\forall \sigma_1, \sigma_2, \sigma_3 \; (\rho(\sigma_1, \sigma_2) \wedge \rho(\sigma_2, \sigma_3)) \supset \rho(\sigma_1, \sigma_3)). \quad (3.1)$$

$$symétrique(\rho) - (\forall \sigma_1, \sigma_2 \; \rho(\sigma_1, \sigma_2) \supset \rho(\sigma_2, \sigma_1)). \quad (3.2)$$

$$antisymétrique(\rho) - (\forall \sigma_1, \sigma_2 \; \rho(\sigma_1, \sigma_2) \supset \neg\rho(\sigma_2, \sigma_1)). \quad (3.3)$$

$$réflexive(\rho) - (\forall \sigma_1 \; \rho(\sigma_1, \sigma_1)). \quad (3.4)$$

$$irréflexive(\rho) - (\forall \sigma_1, \sigma_2 \; \rho(\sigma_1, \sigma_2) \supset \sigma_1 \neq \sigma_2). \quad (3.5)$$

$$injective(\rho) - (\forall \sigma_1, \sigma_2, \sigma_3 \; (\rho(\sigma_2, \sigma_1) \wedge \rho(\sigma_3, \sigma_1)) \supset \sigma_2 = \sigma_3). \quad (3.6)$$

L'équivalence logique « $-$ » introduit ici des conditions nécessaires et suffisantes.

3.1.2 Problèmes de représentation

Nous avons vu à la section précédente la syntaxe générale de $\mathcal{L}$. Nous discutons maintenant comment certains problèmes de représentation sont résolus dans $\mathcal{L}$, à savoir: la représentation de l'incertitude, des poids, des types, des alternatives et de la négation.

a) Certitude

La certitude représente la «*confiance*» qu'on accorde à la véracité d'un fait. Nous ne nous intéressons qu'aux certitudes qui sont *quantifiables*, afin de pouvoir les comparer entre elles. C'est donc dire que de simples annotations de l'encodeur ou du traducteur, du genre «*ce passage n'était pas clair dans le texte original*» ou «*l'orthographe est incorrecte*» sont ignorées.

Voici quelques-unes des sources les plus courantes d'incertitude dans les documents électroniques, et leurs liens avec les types d'information. Des exemples concrets sont donnés plus loin lors de l'introduction des types en question.

- La transcription du texte à partir d'un document original est incertaine. Ceci peut être dû à du bruit sur le signal, ou à un processus de reconnaissance de parole ou de reconnaissance optique de caractères imparfait. Cette incertitude s'applique aux informations de type *contenu-textuel* (voir section 3.2.1);

- L'interprétation sémantique du texte est ambiguë, ce qui peut être dû entre autres à la polysémie. Cette incertitude s'applique au *contenu-sémantique* (voir section 3.2.3);

- La caractérisation du contenu peut ne pas s'appliquer correctement. Par exemple, un mot peut avoir été identifié à tort comme une *abbréviation* par l'encodeur. Cette incertitude s'applique aux informations de type *linguistique* et *logique* (voir sections 3.3.1 et 3.3.2);
- La valeur donnée à un attribut, ou le contenu fourni par l'encodeur est incertain. Il peut s'agir par exemple de l'expansion d'une abbréviation. Cette incertitude s'applique aux *attributs* (voir section 3.3.3);
- Le point précis auquel un élément commence ou se termine est incertain. Par exemple, les limites d'un paragraphe peuvent être floues. Cette incertitude s'applique à la *structure logique* (voir section 3.4.2).

Dans $\mathcal{L}$, lorsque la certitude est omise, elle est supposée égale à $\mu_\top$. Ainsi, la règle suivante est définie:

$$\rho_c(\alpha_1, \alpha_2) - \rho_c(\alpha_1, \alpha_2)[\mu_\top]. \tag{3.7}$$

Un prédicat $\rho_c(\alpha_1, \alpha_2)$de certitude μ implique ce même prédicat avec une certitude μ' pour tout μ' tel que $\mu' \leq \mu$ (règle 3.8). En d'autres termes, un prédicat est vérifié si le même prédicat avec une certitude plus élevée est vérifié.

$$\forall \mu' \; ((\mu' \leq \mu) \wedge \rho_c(\alpha_1, \alpha_2)[\mu]) \supset \rho_c(\alpha_1, \alpha_2)[\mu']. \tag{3.8}$$

b) Poids

Le poids est utilisé lorsque l'on cherche à quantifier le degré ou la force de la relation entre deux éléments. Supposons par exemple le prédicat *thème*, qui identifie le thème d'un document ou de l'un de ses passages. Comme ces thèmes ont une importance relative, c'est-à-dire que certains correspondent mieux à l'idée générale du texte que d'autres, il est utile de leur adjoindre un poids qui mesure cette importance. Ainsi, dans l'exemple ci-dessous, *Mohican* est un thème de s avec un poids de 0.15.

$$thème(s, Mohican)\{.15\}.$$

L'interprétation d'un poids est très similaire à celle d'une certitude – d'où leurs représentations analogues. Les deux règles déjà définies pour les certitudes s'appliquent également aux poids:

$$\rho_p(\alpha_1, \alpha_2) \quad - \quad \rho_p(\alpha_1, \alpha_2)\{\delta_\top\}. \tag{3.9}$$

$$\forall \delta' \; (\delta' \leq \delta) \wedge \rho_p(\alpha_1, \alpha_2)\{\delta\} \quad \supset \quad \rho_p(\alpha_1, \alpha_2)\{\delta'\}. \tag{3.10}$$

c) Types

Les éléments d'information qui contribuent à la définition des documents sont *typés*. Ainsi, le prédicat suivant signifie que s (un passage du document) est de type *chapitre* avec une certitude de .8:

$$chapitre(s)[.8].$$

Les types sont organisés en une hiérarchie, similaire à la grille de types des graphes conceptuels, ou à la subsomption de termes des logiques terminologiques.[7] Cette hiérarchie est définie par des règles, comme la règle ci-dessous qui déclare que *chapitre* est un sous-type de *division*.

$$chapitre(\sigma)[\mu] \supset division(\sigma)[\mu].$$

Dans cet exemple, ainsi que dans toutes les règles de types dans $\mathcal{L}$, la certitude entre types et sous-types est conservée. Pour exprimer le fait qu'un type n'en implique un autre que partiellement, un critère de *distance sémantique* peut être utilisé (voir section 3.3.7).

Les deux types suivants sont définis: le type *universel*, représenté par $t_\top$, et le type *absurde*, représenté par $t_\bot$. Tous les types sont des sous-types de $t_\top$, et le type $t_\bot$ est sous-type de tous les types. Ceci est exprimé par les deux règles suivantes:

$$\forall \rho_t \ \rho_t(\sigma)[\mu] \supset t_\top(\sigma)[\mu]. \tag{3.11}$$

$$t_\bot(\sigma)[\mu] \supset \forall \rho_t \ \rho_t(\sigma)[\mu]. \tag{3.12}$$

Les types ne sont pas mutuellement exclusifs, à moins qu'il n'en soit spécifié autrement par une règle. Par exemple, il est tout à fait possible d'avoir un segment de texte qui est à la fois *abbréviation* et *terme-technique*.

Notons enfin que par la suite, plutôt que de donner les règles d'héritage des types comme ci-dessus, nous préférons présenter une arborescence graphique (voir par exemple la figure 3.1). Un lien $a \longrightarrow b$ signifie que b est un sous-type de a, ce qui s'écrit aussi par la règle $b(\sigma)[\mu] \supset a(\sigma)[\mu]$.

d) Alternatives

Les alternatives surviennent lorsqu'un élément peut être exprimé de diverses façons. En plus des sources déjà citées pour les certitudes, les alternatives peuvent être exprimées

7. Tout comme dans ces langages, la relation de *type* définit une ordre partiel, c'est-à-dire qu'elle est réflexive, anti-symétrique et transitive. Ces contraintes ne sont pas définies ici puisque les types sont décrits exhaustivement.

explicitement dans le texte pour le «*ou*» ou par d'autres opérateurs discursifs. Un exemple typique d'alternative est la polysémie, qui survient quand un élément du langage a plusieurs sens possibles.

Les différentes alternatives d'un élément d'information sont liées à cet élément par le biais du `Prédicat-P` *alternative*. Ainsi, en supposant un objet s, les prédicats ci-dessous déclarent deux alternatives pour cet objet, s_1 et s_2, ayant des poids respectifs de .7 et .3:

$$\begin{array}{l} \mathit{alternative}(s, s_1)\{.7\}. \\ \mathit{alternative}(s, s_2)\{.3\}. \end{array}$$

Le poids qui est associé à une alternative est toujours pris au sens individuel, c'est-à-dire qu'il s'applique à une alternative indépendamment des autres. Cela signifie entre autre que la somme des poids des alternatives pour un élément donné n'est pas nécessairement égale à $\delta_\top$.

Les faits déclarés pour s sont considérés comme des caractéristiques communes à s_1 et s_2, et sont donc hérités par ces derniers. Tout ce qui distingue s_1 de s_2 est déclaré comme un fait pour l'objet en question. Soit par exemple:

$$\begin{array}{l} \mathit{contenu\text{-}textuel}(s). \\ \mathit{alternative}(s, s_1). \\ \mathit{alternative}(s, s_2). \\ \mathit{chapitre}(s_1). \end{array}$$

Ici, s_1 et s_2 partagent le fait d'être de type *contenu-textuel*; s_1 est en plus de type *chapitre*.

De façon plus générale, soit un objet α pour lequel un fait ρ est défini; alors ce fait s'applique aussi à toutes les alternatives α' de α. Ceci est vrai autant pour les `Prédicat-C` que pour les `Prédicat-P` (à l'exception bien sûr du prédicat *alternative* lui-même). Dans les deux cas, la certitude ou le poids sur la propriété de α' est héritée de α, et ce quel que soit le poids de l'alternative. Les poids sur les alternatives ne peuvent influer que sur les faits qui leur sont propres.

Cet héritage des alternatives est formalisé par les règles suivantes:

$$\forall\alpha' \; (\rho_c(\alpha, \beta)[\mu] \wedge \mathit{alternative}(\alpha, \alpha')\{\delta\}) \;\supset\; \rho_c(\alpha', \beta)[\mu]. \tag{3.13}$$

$$\forall\alpha' \; (\rho_p(\alpha, \beta)\{\delta\} \wedge \mathit{alternative}(\alpha, \alpha')\{\delta'\} \wedge (\rho_p \neq \mathit{alternative})) \;\supset\; \rho_p(\alpha', \beta)\{\delta\}. \tag{3.14}$$

e) Négation

Le négation pose un problème complexe pour la représentation de connaissances; si son apport est important du point de vue de l'expressivité, elle peut entraîner des problèmes de cohérence dans le langage de représentation [Par94b]. Puisqu'il s'agit dans notre cas de représentation de textes, elle partage l'ambiguïté de la négation dans la langue naturelle, qui peut prendre plusieurs sens [Car83, pp187–194] selon l'effet ou la fonction désirée. Le problème est donc complexe, et nous ne prétendons pas lui apporter une solution complète: nous nous contentons de décrire une utilisation restreinte de la négation dans notre modèle.

Nous distinguons la négation *linguistique*, exprimée par des opérateurs discursifs tels que «*ne...pas*», de la négation *logique*, exprimée par des connecteurs logiques. Nous ne proposons pas d'expression pour la négation linguistique dans notre modèle. Les opérateurs discursifs n'ayant pas d'interprétation logique simple, il est difficile de bien les représenter dans $\mathcal{L}$. De plus, l'intérêt de savoir traiter ce type de négation pour répondre à des requêtes thématiques n'est pas évident. Si l'on admet que la fonction principale de la négation est «*la correction d'une préconception antérieure*» [Car83, p188], alors les deux énoncés – l'affirmation et la négation – pourraient bien être tous deux pertinents.

Sur le plan *logique*, nous distinguons deux grandes fonctions de la négation en recherche d'informations:[8] soit pour déclarer l'orthogonalité entre deux concepts, soit pour exprimer la non-appartenance d'un concept aux index d'un passage.

L'exclusion mutuelle – ou l'orthogonalité – de deux concepts est exprimée au niveau des connaissances par le biais d'un prédicat intermédiaire. Par exemple, pour affirmer qu'un élément de type *pingouin* ne peut pas être aussi de type *vole*, on définit la *règle* suivante:

$$pingouin(\alpha) \supset \neg vole(\alpha).$$

L'appartenance ou la non-appartenance d'un index à un passage est du ressort des *faits*. Or, tout comme ils ne peuvent admettre une certitude ou un poids nul, les faits ne peuvent être qu'affirmatifs. Ce type de négation est donc implicitement représenté par l'hypothèse de monde fermé: l'absence de $thème(\sigma, \varsigma)$ dans la base de faits est logiquement équivalente à $\neg thème(\sigma, \varsigma)$.

3.1.3 Interprétation de $\mathcal{L}$

Nous avons défini dans les sections précédentes la syntaxe de $\mathcal{L}$. Cette définition devrait à elle seule suffire à comprendre la suite de notre travail, puisque nous sommes essentiellement intéressés par l'aspect représentationnel de $\mathcal{L}$. Toutefois, afin de lever toute ambiguïté quant à la sémantique de $\mathcal{L}$, nous donnons ici sa *fonction d'interprétation*.

La sémantique de $\mathcal{L}$ est inspirée de KIF [GF92], un formalisme approprié pour la définition d'ontologies. Toutefois, un ajout important de $\mathcal{L}$ par rapport à KIF est qu'il permet

8. Outre la déclaration de contraintes ou de règles du langage de représentation.

l'expression de certitudes, poids et interprétations multiples sur les items d'information. Une autre différence concerne la quantification universelle, qui est prise par défaut pour les variables libres dans $\mathcal{L}$.

Soit la *fonction d'affectation* v, qui assigne une constante à chaque variable ou constante de $\mathcal{L}$:

$$v : \mathcal{F}_{vc} \longrightarrow \mathcal{F}_c$$

où $\mathcal{F}_{vc}$ est l'ensemble des variables et constantes, et $\mathcal{F}_c$ est l'ensemble des constantes. Pour les constantes, $\mathcal{F}_{vc}$ est la fonction d'identité.

La *fonction d'interprétation* i pour une affectation de variables v, assigne à chaque formule de $\mathcal{L}$ une valeur de vérité:

$$i_v : \mathcal{F}_\phi \longrightarrow \{vrai, faux\}$$

où $\mathcal{F}_\phi$ est l'ensemble des `Formules` de $\mathcal{L}$.

L'interprétation de constantes ou de variables seules n'a pas de sens ici, puisque la fonction d'interprétation est limitée à des valeurs booléennes. Nous ne définissons pas non plus d'*interprétation* de la certitude ou du poids, au sens de modèle probabiliste, ou de croyance, confiance, etc. Cette tâche ne prend de sens que lors de la combinaison de ces mesures entre elles, et ne peut donc être définie que lors du calcul de la pertinence.

Soit $\mathcal{F}_\rho$ l'ensemble des faits, qui sont soit déclarés, soit déduits à partir de règles. On dit alors qu'un `Prédicat-C` de certitude μ est *vrai* dans $\mathcal{L}$ s'il existe dans $\mathcal{F}_\rho$ avec une certitude μ' supérieure ou égale à μ.

$$i_v(\rho_c(\alpha_1)[\mu]) = \begin{cases} vrai & \text{si } v(\rho_c)(v(\alpha_1))[v(\mu')] \in \mathcal{F}_\rho \text{ pour } \mu \leq \mu' \\ faux & \text{sinon} \end{cases}$$

$$i_v(\rho_c(\alpha_1, \alpha_2)[\mu]) = \begin{cases} vrai & \text{si } v(\rho_c)(v(\alpha_1), v(\alpha_2))[v(\mu')] \in \mathcal{F}_\rho \text{ pour } \mu \leq \mu' \\ faux & \text{sinon} \end{cases}$$

La relation « $\leq$ » utilisée ici définit un ordre total sur les certitudes, c'est-à-dire que pour tout couple de certitudes, elle est réflexive, anti-symétrique et transitive. La même règle vaut pour les `Prédicat-P`:

$$i_v(\rho_p(\alpha_1, \alpha_2)\{\delta\}) = \begin{cases} vrai & \text{si } v(\rho_p)(v(\alpha_1), v(\alpha_2))[v(\delta')] \in \mathcal{F}_\rho \text{ pour } \delta \leq \delta' \\ faux & \text{sinon} \end{cases}$$

La certitude et le poids sont maximaux par défaut:

$$i_v(\rho_c(\alpha_1, \alpha_2)) = i_v(\rho_c(\alpha_1, \alpha_2)[\mu_\tau]).$$
$$i_v(\rho_p(\alpha_1, \alpha_2)) = i_v(\rho_p(\alpha_1, \alpha_2)\{\delta_\tau\}).$$

Ces quatre dernières définitions correspondent aux règles 3.7 à 3.10.

La négation est définie par:

$$i_v(\neg\phi) = \begin{cases} vrai & \text{si } i_v(\phi) = faux \\ faux & \text{sinon} \end{cases}$$

La conjonction est définie comme:

$$i_v(\phi_1 \wedge \phi_2) = \begin{cases} vrai & \text{si } i_v(\phi_1) = vrai \text{ et } i_v(\phi_2) = vrai \\ faux & \text{sinon} \end{cases}$$

La disjonction, l'implication matérielle et l'équivalence logique sont définies à l'aide de la négation:

$$\begin{aligned} i_v(\phi_1 \vee \phi_2) &= i_v(\neg(\neg\phi_1 \wedge \neg\phi_2)) \\ i_v(\phi_1 \supset \phi_2) &= i_v(\neg\phi_1 \vee \phi_2)) \\ i_v(\phi_1 - \phi_2) &= i_v((\phi_1 \wedge \phi_2) \vee (\neg\phi_1 \wedge \neg\phi_2)) \end{aligned}$$

L'égalité dans $\mathcal{L}$ est restreinte à des variables ou constantes:

$$i_v(\alpha_1 = \alpha_2) = \begin{cases} vrai & \text{si } v(\alpha_1) = v(\alpha_2) \\ faux & \text{sinon} \end{cases}$$

La non-égalité est définie par le biais de la négation:

$$i_v(\alpha_1 \neq \alpha_2) = i_v(\neg(\alpha_1 = \alpha_2))$$

Quand à l'inégalité, elle est définie pour les certitudes et pour les poids comme suit:

$$i_v(\mu \leq \mu') = \begin{cases} vrai & \text{si } \mu \leq \mu' \\ faux & \text{sinon} \end{cases}$$

$$i_v(\delta \leq \delta') = \begin{cases} vrai & \text{si } \delta \leq \delta' \\ faux & \text{sinon} \end{cases}$$

Le quantificateur existentiel est donné par:

$$i_v(\exists\nu_1, \ldots, \nu_k\phi) = \begin{cases} vrai & \text{si } \exists v' \, i_{v'}(\phi) = vrai \\ faux & \text{sinon} \end{cases}$$

où v' représente une fonction d'affectation qui diffère de v seulement par rapport aux variables $\nu_1, \ldots, \nu_k$. En d'autres termes, la formule ϕ est vérifiée dans v si il est possible de modifier l'affectation des variables $\nu_1, \ldots, \nu_k$ de façon à ce que ϕ soit vérifiée (cette nouvelle fonction d'affectation étant v').

Le quantificateur universel est défini de la même façon:

$$i_v(\forall \nu_1, \ldots, \nu_k \phi) = \begin{cases} vrai & \text{si } \forall v' \, i_{v'}(\phi) = vrai \\ faux & \text{sinon} \end{cases}$$

Enfin, la fonction d'interprétation i donne la valeur de vérité d'une formule ou d'une règle ϕ quelle que soit la fonction d'affectation. $i(\phi)$ est vérifiée si $i_v(\phi)$ est vraie pour toute fonction d'affectation v.

$$i(\phi) = \begin{cases} vrai & \text{si } \forall v \, i_v(\phi) = vrai \\ faux & \text{sinon} \end{cases}$$

Ceci traduit le fait que les variables libres sont quantifiées universellement par défaut.

3.2 Représentation du contenu

Après ce bref aperçu du langage, nous attaquons maintenant la descriptions exhaustive des éléments du langage. Nous présentons dans cette section informations de type *contenu*. Les informations de *méta-contenu* et de *structure de contenu* sont décrites dans les sections suivantes.

La notion de contenu joue un rôle prépondérant dans notre modèle puisque tous les autres types d'informations s'organisent autour d'elle. Reprenant la dichotomie signifiant/signifié, nous distinguons les symboles (ou le *contenu-signifiant*, qui peuvent encore êtres divisés en contenu *textuel* ou *non-textuel*) de leur signification (*contenu-sémantique*). Ces types sont représentés à la figure 3.1. Les triangles sous *contenu-textuel*, *contenu-non-textuel* et *contenu-sémantique* représentent des sous-hiérarchies qui sont décrites dans les sections correspondantes ci-dessous.

3.2.1 Contenu textuel

Le contenu textuel d'un document représente tout ce qui apparaît dans le document imprimé ou visualisé. Il s'agit bien entendu du corps du texte, mais aussi de toute information qui le complète ou le reformule, comme les titres, les notes de bas de page, etc. Le contenu supplémentaire fourni par l'encodeur ou le dispositif automatique d'encodage fait

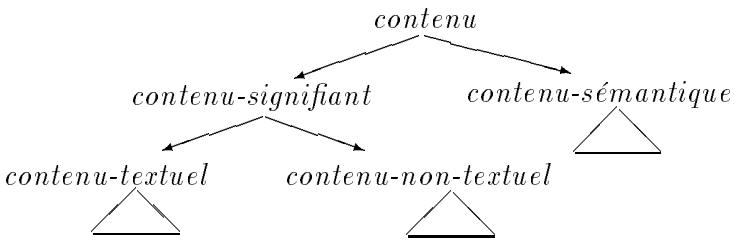


Figure 3.1. Types de contenu

quant à lui partie du *méta-contenu*, à l'exception des *corrections*, qui sont vues ici comme fournissant une autre interprétation du texte originel.

Le texte est généralement considéré par les applications informatiques comme une chaîne de caractères. Les humains ne procèdent pas de cette manière: ainsi, la plupart des lecteurs adultes peuvent lire et comprendre un mot comme un tout sans avoir à le décoder lettre par lettre. De même, la décomposition en lettres est généralement de peu d'utilité en recherche d'informations, si ce n'est pour la phase de lemmatisation. Par contre, une telle décomposition peut être utile à un niveau plus élevé, comme par exemple en permettant la combinaison de mots pour former les syntagmes.

Notre approche consiste donc à considérer les chaînes de caractères comme atomiques ou indivisibles, mais de permettre leur combinaison à un niveau plus élevé. Nous nous employons surtout dans cette section à décrire la représentation des éléments atomiques, c'est-à-dire formés d'une seule chaîne de caractères. La combinaison d'éléments afin de former les passages du texte sera discutée à la section 3.4.

Les chaînes de caractères sont des constantes dans $\mathcal{L}$. Le prédicat *chaîne* est vérifié pour toutes les chaînes, c'est-à-dire que pour toute chaîne τ, on a $\mathit{chaîne}(\tau)$.

Les passages textuels du document sont représentés par des éléments de type *contenu-textuel*. Les passages ainsi représentés peuvent aller du simple mot au document entier. S'il s'agit d'éléments atomiques, on peut leur associer une *chaîne* par la relation *texte*, dont le premier argument est un élément de type *contenu-textuel*, et le second, une chaîne de caractères (règle 3.15).

$$\mathit{texte}(\sigma, \tau) \supset (\mathit{contenu\text{-}textuel}(\sigma) \wedge \mathit{chaîne}(\tau)). \tag{3.15}$$

Cette représentation permet de distinguer les chaînes de caractères de leurs occurences dans le texte. De cette façon, une même chaîne de caractères peut apparaître plusieurs fois dans le texte. Par exemple, la chaîne «*Mohicans*» peut être associée aux passages s_1 et s_2 comme suit:

contenu-textuel(s_1).
texte(s_1, «*Mohicans*»).
contenu-textuel(s_2).
texte(s_2, «*Mohicans*»).

Les incertitudes résultent habituellement de la conversion du document originel en sa forme électronique: il est possible qu'une partie ne soit pas imprimée clairement, ou que des processsus comme la reconnaissance de caractères ou de parole ne soient pas fiables. Les certitudes sur le contenu textuel sont toujours reportées sur la relation *texte*. Ainsi, dans l'exemple ci-dessous, le texte associé à s_1 a une certitude de .8.

contenu-textuel(s_1).
texte(s_1, «*Mohicans*»)[.8].

Dans l'exemple ci-dessous, les deux *alternatives* sont «*Mohicans*» et «*Mojicans*». Les mesures accompagnant les prédicats *alternative*s sont des poids, qui signifient dans le cas présent que «*Mohicans*» est 4 fois plus probable que «*Mojicans*».

contenu-textuel(s).
alternative(s, s_1){.8}.
alternative(s, s_2){.2}.
texte(s_1, «*Mohicans*»).
texte(s_2, «*Mojicans*»).

À noter la différence entre cette représentation et la représentation ci-dessous, où aucun poids n'est donné pour les alternatives mais une certitude est associée à chaque *texte*. Ces deux représentations ne sont pas équivalentes.

contenu-textuel(s).
alternative(s, s_1).
alternative(s, s_2).
texte(s_1, «*Mohicans*»)[.8].
texte(s_2, «*Mojicans*»)[.2].

Les *corrections* sont des rectifications au texte originel apportées par l'encodeur ou le processus d'encodage. Elles sont considérées dans notre modèle comme des *alternatives*, où le texte originel a un poids supérieur à celui de la correction. Ainsi, dans l'exemple ci-dessous, «*Mohicans*» est substitué au texte originel «*Mojicans*».

contenu-textuel(s).
alternative(s, s_1){.9}.
alternative(s, s_2){.1}.
texte(s_1, «*Mohicans*»).
texte(s_2, «*Mojicans*»).

3.2.2 Contenu non-textuel

Les éléments non-textuels, tels que les images ou figures, sont représentés à l'aide d'un objet servant de lien vers une entité externe. Les éléments graphiques qui ne contribuent pas au discours, comme par exemple, dans ce travail, les lignes qui précèdent et suivent chaque figure, sont ignorés.

Ainsi, dans l'exemple ci-dessous, l'élément *fig1* est associé au fichier «*fig1.bmp*», qui correspond à une image *bitmap*:

contenu-non-textuel(*fig1*).
externe(*fig1*, «*fig1.bmp*»).

Une mesure de certitude peut aussi accompagner les informations non-textuelles; elle s'exprime alors sur la propriété *externe*. Pour une image, cette certitude traduit l'imprécision du processus de digitalisation, ou la perte d'information due à un certain type d'encodage, comme par exemple à l'imprécision inhérente au format «*jpeg*».

3.2.3 Contenu sémantique

De la même façon que nous n'avons pas spécifié les détails de la représentation interne des chaînes de caractères et des éléments non-textuels, considérant pour les uns qu'il s'agit de constantes atomiques, et se référant pour les autres à des entités externes, nous ne définirons pas ici de sous-langage pour la représentation du contenu sémantique. Une telle tâche serait de toute façon dépendante du domaine d'application. Toutefois, trois problèmes communs à toutes les représentations peuvent être exprimés ici, à savoir: l'incertitude, la polysémie et le paraphrasage [Par94b].

Le contenu sémantique est représenté par le type *contenu-sémantique*. Un élément du signifiant est lié à un contenu sémantique par la relation *sém*, dont le premier argument est un élément de type *contenu-signifiant*, et le second, de type *contenu-sémantique* (règle 3.16).

$$sém(\sigma, \varsigma) \supset (contenu\text{-}signifiant(\sigma) \wedge contenu\text{-}sémantique(\varsigma)). \quad (3.16)$$

Ainsi, dans l'exemple ci-dessous, s a pour représentation sémantique l'élément *mohican*. Ce dernier peut représenter un mot-clé, concept, graphe, etc. Il ne doit pas être confondu avec la chaîne «*Mohican*»; il ne s'agit que d'un identificateur qui renvoie à une information sémantique, mais qui aurait tout aussi bien pu s'appeler *m42*.

contenu-sémantique(*mohican*).
sém(*s*, *mohican*).

La certitude du contenu sémantique s'exprime sur le prédicat *sém*; en fait ce qu'on exprime ici, ce n'est pas tant la certitude du contenu sémantique en lui-même, mais bien la certitude qu'il s'applique au contenu textuel auquel il est lié.

Le fait qu'une représentation sémantique soit liée à un contenu textuel n'implique pas une bijection: ainsi, les problèmes de polysémie et paraphrasage peuvent survenir.

Le paraphrasage se traduit en langue naturelle par l'emploi de synonymes, périphrases, métaphores, etc. Une tendance consiste à représenter les énoncés «*équivalents*» sous une même forme, ce qui évite d'avoir un mécanisme d'inférence pour mettre en équivalence les connaissances lors de la correspondance. Le problème avec cette approche est la perte d'information qu'elle entraîne. Ainsi, même si l'on peut dire que la «*ville lumière*» est quasi-synonyme de «*Paris*», cette expression ne véhicule pas le même sens ou les mêmes connotations. Woods parle du «*mythe de la forme canonique*» pour désigner ce phénomène [Woo75]. Il remarque d'abord qu'il est probablement impossible de résoudre, c'est-à-dire de réduire à leur forme standard, toutes les paraphrases d'une langue. Mais surtout, l'implication entre deux paraphrases est souvent unidirectionelle: une expression en implique une autre mais pas le contraire. Il faudra donc de toute manière un mécanisme d'inférence lors de la correspondance.

Quoi qu'il en soit, le paraphrasage peut être représenté en faisant pointer deux objets de contenu textuel sur le même objet de contenu sémantique. Ainsi, dans l'exemple ci-dessous, les chaînes «*Paris*» et «*la ville lumière*» réfèrent toutes deux à *paris*:

texte(s_1, «*paris*»).
texte(s_2, «*la ville lumière*»).
sém(s_1, *paris*).
sém(s_2, *paris*).
contenu-sémantique(*paris*).

Le problème de la polysémie est l'inverse du paraphrasage: elle survient quand un élément de contenu textuel a plusieurs sens possibles. De même que pour le paraphrasage, il est montré dans [Kay84] qu'il n'est pas toujours avantageux de résoudre ces ambiguïtés dans la représentation. Ainsi, si la requête est délibérément ambiguë, ou si la résolution de cette ambiguïté n'apporte rien au processus de recherche, il vaut mieux l'intégrer dans la

représentation plutôt que de chercher à la résoudre.

La polysémie est représentée par des alternatives. Ainsi, l'objet sémantique *remercier* peut avoir les deux interprétations suivantes:

contenu-sémantique(*remercier*).
alternative(*remercier*, *savoir-gré*).
alternative(*remercier*, *congédier*).

3.3 Représentation du méta-contenu

Le *méta-contenu* représente l'information *à-propos* du contenu, c'est-à-dire l'information qui se rattache au contenu en précisant sa nature ou en y ajoutant des informations telles que les attributs de contenu, les thèmes, les connaissances, etc.

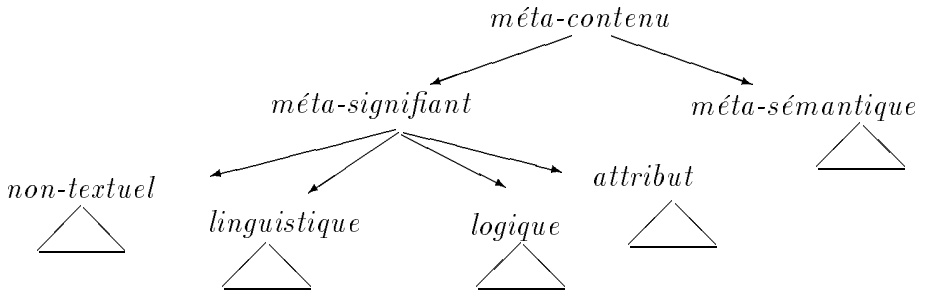


Figure 3.2. Types de méta-contenu

La figure 3.2 résume les différents types de méta-contenu. On distingue deux grands sous-types: le type *méta-signifiant*, ou les informations qui servent à caractériser le *contenu-signifiant*, et le type *méta-sémantique*, pour la caractérisation du *contenu-sémantique*. Ces types peuvent aussi être vus comme une information additionnelle sur le contenu, et sont donc définis comme des sous-types des types de contenu correspondants. On a donc:

$$\textit{linguistique}(\sigma) \supset \textit{contenu-textuel}(\sigma). \tag{3.17}$$
$$\textit{logique}(\sigma) \supset \textit{contenu-textuel}(\sigma). \tag{3.18}$$
$$\textit{non-textuel}(\sigma) \supset \textit{contenu-non-textuel}(\sigma). \tag{3.19}$$
$$\textit{méta-sémantique}(\sigma) \supset \textit{contenu-sémantique}(\sigma). \tag{3.20}$$

Nous détaillons d'abord dans les sections qui suivent les informations de type *linguistique* et *logique*. Les attributs qui sont liés à ces types sont introduits au fur et à mesure

qu'ils sont requis. Les autres types *attribut* sont ensuite présentés. Nous discutons enfin de la caractérisation des informations non-textuelles, de contenu sémantique, de thèmes et de connaissances.

3.3.1 Linguistique

Les informations de l'ordre de la linguistique représentent le texte que l'auteur ou l'encodeur cherche à distinguer du texte qui l'entoure pour des raisons diverses. Elles se traduisent souvent dans le document imprimé par des styles ou des procédés typographiques divers: comme par exemple l'italique, les caractères gras, guillemets, etc. Nous ne cherchons pas ici à identifier ces styles, mais bien les raisons qui motivent leur utilisation. Lorsque ces raisons sont inconnues, cette information est tout simplement ignorée.

Tous les éléments de type *linguistique* sont forcément aussi de type *contenu-textuel* (règle 3.17). Puisque cette information peut être déduite à partir des règles de typage, elle est omise des représentations. Ceci vaut pour tous les types.

La figure 3.3 résume les différents types d'information *linguistique*. On distingue deux grands types, *carac-forme* et *carac-sens*, qui identifient respectivement des éléments dont la *forme* ou le *sens* sont caractérisés. Nous voyons maintenant en détails comment ces types s'expriment dans les documents et dans $\mathcal{L}$.

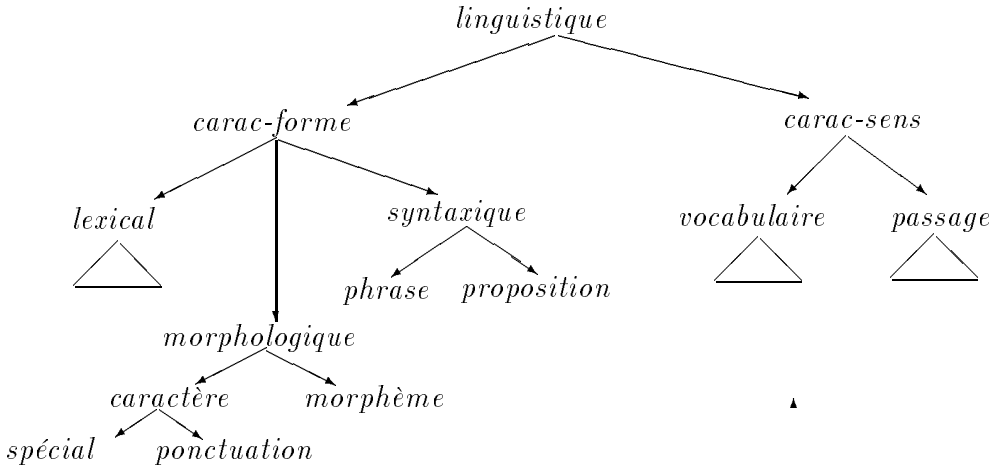


Figure 3.3. Types de contenu linguistique

a) Informations morphologiques

Les types *morphème* et *caractère* sont des annotations morphologiques: le *morphème* est associé ici au lexème, ou morphème lexical, et est surtout inclus par souci de flexibilité, son usage pour la recherche d'informations étant surtout limité à la phase de *lemmatisation*. Il en va de même pour le type *caractère*; plutôt que décomposer chaque mot en caractères, nous utilisons ce type pour identifier la *ponctuation* ou certains symboles spéciaux qui ne pourraient pas être identifiés autrement.

b) Informations lexicales

Le type *lexical* permet d'identifier la catégorie lexicale d'un mot. Ces types sont définis à la figure 3.4. Cette liste ne représente bien entendu qu'un sous-ensemble des catégories lexicales; ne sont retenues ici que celles qui peuvent aider à l'identification des thèmes.

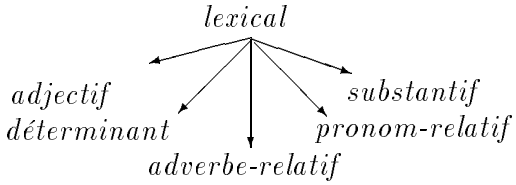


Figure 3.4. Types de contenu lexical

- le type *adjectif* désigne les adjectifs qualificatifs – qu'ils soient ou non épithètes – ou les locution adjectivales;
- on regroupe sous le vocable de *déterminant*, les articles et les adjectifs non-qualificatifs, comme «*le*», «*cet*», etc.;
- les *adverbe-relatif*s, comme «*où*»;
- les *pronom-relatif*s joignent un nom ou un pronom à une proposition subordonnée relative. Il s'agit de «*qui*», «*que*», «*quoi*», «*lequel*», etc.;
- les *substantif*s ou noms.

c) Informations syntaxiques

Les annotations syntaxiques identifient le rôle fonctionnel du texte par rapport au texte qui l'entoure. Il peut s'agir d'une *proposition* ou d'une *phrase*. Cette information peut bien sûr être utile pour une analyse syntaxico-sémantique lors de l'indexation, mais également pour une recherche en texte intégral.

d) Vocabulaire

Le type *vocabulaire* regroupe les items qui s'adressent au vocabulaire ou à la formation d'éléments du vocabulaire à partir d'éléments plus simples comme les morphèmes ou les caractères. Ces types sont définis à la figure 3.5.

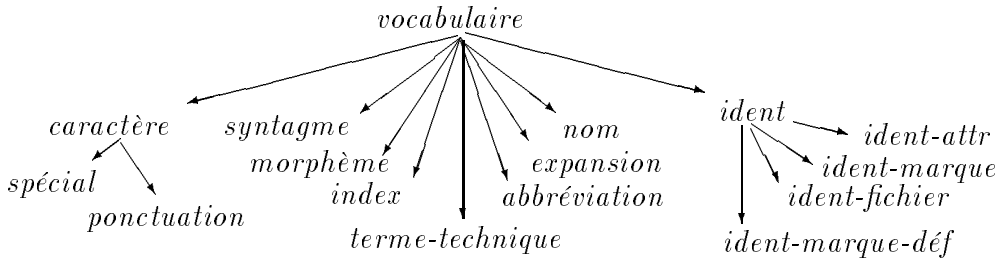


Figure 3.5. Types de vocabulaire

– Le type *syntagme* est utile en recherche d'informations, puisqu'il permet d'identifier les éléments du vocabulaire qui sont *composés*, en particulier les groupes nominaux. L'exemple ci-dessous représente le groupe nominal «*recherche d'informations*»:

syntagme(*s*).
texte(*s*, «*recherche d'informations*»).

– le type *terme-technique* est utilisé pour marquer un mot simple ou un groupe de mots considérés comme un *terme technique*. Ces mots apparaissent souvent en italique ou en caractère gras dans le texte lors de leur première mention. Par exemple:

terme-technique(*s*).
texte(*s*, «*mémoire vive*»).

– le type *abbréviation* est utilisé pour les abbréviations comme les acronymes, contractions, noms d'organismes, etc. On peut associer une *expansion* à l'abbréviation, à l'aide du prédicat *expan*. Ainsi, dans l'exemple ci-dessous, l'encodeur a précisé que l'abbréviation «*RI*», désignée ici par s_1, avait pour expansion «*recherche d'informations*», désignée par s_2, par la relation $expan(s_1, s_2)$.

abbréviation(s_1).
texte(s_1, «*RI*»).
expan(s_1, s_2).
expansion(s_2).
texte(s_2, «*recherche d'informations*»).

Le type *expansion* est symétrique à *abbréviation*, c'est-à-dire qu'il existe une relation *abbr* qui est l'inverse de *expan*:

$$expan(\sigma_1, \sigma_2)[\mu] - (abbr(\sigma_2, \sigma_1)[\mu'] \wedge \mu' \geq \mu). \tag{3.21}$$

La certitude de la relation d'abbréviation est supérieure ou égale à la relation d'expansion à cause de la polysémie des expansions. Par exemple, «*IBM* » peut être l'abbréviation de «*International Business Machines*» ou de «*Inter-Ballistic Missiles*».

- le type *nom* réfère au nom d'un individu, organisme, établissement, région géographique, etc. Par exemple, en parlant de la gare St-Lazare, on peut avoir:

  nom(s).
  texte(s, «*St-Lazare*»).

- le type *index* sert à marquer une expression pour la *table d'index* du document. La table d'index est une table alphabétique des sujets ou des noms cités dans le document, accompagnés de leur référence. Ces marques sont généralement placées dans les documents sources afin de permettre la génération automatique de la table d'index; dans $\mathcal{L}$, elles nous permettent d'identifier facilement un thème.

- le type *ident* désigne un *identificateur*, c'est-à-dire le nom d'un fichier, d'un élément SGML, etc. Ainsi, dans cet ouvrage, tous les noms de prédicats peuvent être considérés comme des *ident*s. Les sous-types suivants sont définis: *ident-fichier* (nom de fichier), *ident-marque* (identificateur SGML), *ident-marque-déf* (définition d'un identificateur SGML), et *ident-attr* (attribut SGML).

e) Passages

Le type *passage* permet la caractérisation du texte d'une granularité plus grande que le *vocabulaire*; il ne s'agit donc plus de mots mais bien de passages du texte. Les types de *passage* sont donnés à la figure 3.6.

- *étranger*. Identifie un mot ou un groupe de mots appartenant à une langue autre que celle du texte principal. Cette information est cruciale dans un corpus multilingue,

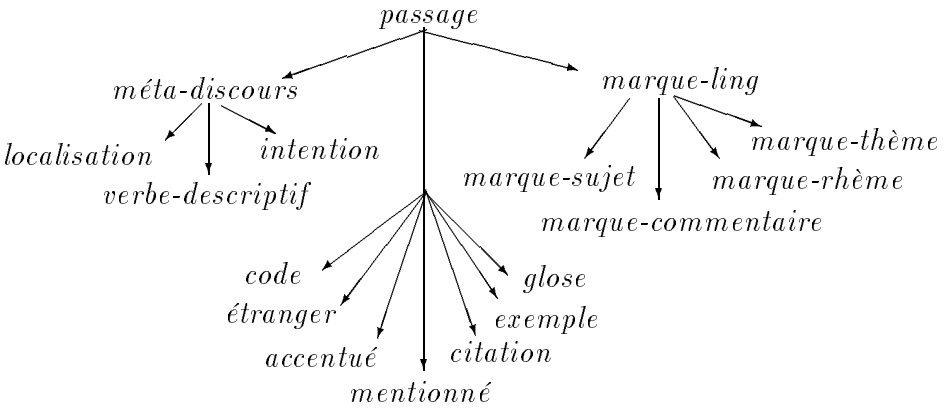


Figure 3.6. Types de passages

puisque les techniques d'indexation varient selon la langue. Ainsi, dans l'exemple ci-dessous, s est un passage du texte écrit en anglais.

étranger(s).
texte$(s,$ «*frame*»$)$.
langue$(s,$ *anglais*$)$.

On peut spécifier la langue dont il s'agit à l'aide du prédicat *langue*, comme dans l'exemple ci-dessus, ou de l'alphabet utilisé à l'aide de la propriété *alphabet*. Ces deux propriétés peuvent en fait être utilisées avec tout *contenu-textuel*, y compris le document lui-même.

– *accentué*. Identifie des mots ou groupes de mots qui sont *accentués* dans le texte, c'est-à-dire qui sont mis en évidence pour des raisons de rhétorique ou de linguistique. Considérons par exemple l'énoncé suivant: «cette station de ski a aussi une vocation *estivale*». Dans cet exemple, «*estivale* » est accentué afin d'augmenter son impact dans le discours en l'opposant implicitement aux activités *hivernales* de la station.

accentué(s).
texte$(s,$ «*estivales*»$)$.

Les informations accentuées peuvent généralement être considérées comme de bon indicateurs du «*centre d'intérêt*» (commentaire) d'une phrase.

– *mentionné*. Identifie des mots ou groupes de mots qui sont *mentionnés* mais pas *utilisés* dans le texte, c'est-à-dire qui contribuent au discours au niveau du signifiant

mais pas du signifié. Ainsi, dans cet ouvrage, le texte placé entre guillemets typographiques (« ») est vu comme un exemple linguistique, c'est dire qu'on s'intéresse aux symboles eux-mêmes plutôt qu'à leur sémantique.

- Le type *glose* désigne une annotation entre les lignes ou en marge d'un texte, pour expliquer un mot difficile, ou éclaircir un passage obscur. Ce type est souvent utilisé en conjonction avec le type *terme-technique*.

- *citation*. Contient un passage qui est attribué par le narrateur ou l'auteur à quelqu'un d'externe au texte. Dans le cas d'un dialogue, il peut s'agir d'un personnage qui parle. Dans un dictionnaire, cela peut être un exemple d'usage. Enfin, il peut aussi s'agir d'un extrait d'un autre document. Par exemple, la citation dénotée par s ci-dessous est attribuée à Louis XIV par la relation *auteur-citation*:

  $citation(s)$.
  $texte(s, \text{«}L'état\ c'est\ moi\text{»})$.
  $auteur\text{-}citation(s, \text{«}LouisXIV\text{»})$.

- Le type *exemple* est une assertion qui sert à illustrer le propos du texte. Tous les propositions linguistiques du chapitre précédent ou les exemples d'utilisation de $\mathcal{L}$ dans ce chapitre pourraient être de ce type.

- Le type *code* est utilisé pour identifier un passage qui est exprimé dans un langage formel, habituellement informatique ou mathématique. Par exemple, dans cet ouvrage, les éléments du langage $\mathcal{L}$, qu'ils apparaissent dans des exemples ou qu'ils soient cités dans le texte, sont de ce type.

- Le type *marque-ling* permet d'identifier le thème d'une phrase, selon les critères linguistiques présentés au chapitre précédent. Nous retenons deux types de thèmes: le *thème* de l'analyse statutaire (*marque-thème*), et le *sujet* des travaux TFA (*marque-sujet*). En conjonction avec ces types il est aussi possible d'identifier le *rhème* (*marque-rhème*) et le *commentaire* (*marque-commentaire*). Par exemple, on peut identifier le thème et le rhème de l'énoncé «*Il mange du poisson*» de la façon suivante:

  $marque\text{-}thème(s_1)$.
  $marque\text{-}rhème(s_2)$.
  $texte(s_1, \text{«}Il\text{»})$.
  $texte(s_2, \text{«}mange\ du\ poisson\text{»})$.

  Seules les informations textuelles sont identifiées de cette façon. Ainsi, dans l'exemple ci-dessus, la marque *présent* qui fait partie du thème, et *affirmatif*, *indicatif* qui font partie du rhème, ne sont pas représentées.

– On peut finalement identifier certaines expressions-outils à l'aide du type *méta-discours*. Cette information est utile afin de déterminer les *intentions* de discours, qui à leur tour indiquent les thèmes. On distingue les expressions de type *verbe-descriptif* et de *localisation*; les premières dénotent des expressions formées autour de verbes *descriptifs* tels que «*décrire*», «*présenter*», etc., alors que les secondes dénotent un ouvrage ou une partie d'un ouvrage, comme par exemple: «*cet article*», «*le chapitre 2*», etc. Par exemple:

verbe-descriptif(s_1).
localisation(s_2).
texte(s_1, «*Nous décrivons*»).
texte(s_2, «*dans ce chapitre*»).

On peut associer une *localisation* à l'élément qui est dénoté par sa chaîne de caractères à l'aide du prédicat *désigne*. Pour l'exemple ci-dessus, en supposant que le chapitre courant soit *chap*3, on aurait:

désigne(s_2, *chap*3).

f) Incertitude linguistique

Les éléments dont nous avons discuté dans cette section peuvent avoir un degré de certitude associé, lorsque l'encodeur ou le processus d'encodage ne parviennent pas à identifier de façon sûre le type de l'élément.

Dans l'exemple ci-dessous, «*recherche d'informations*» est identifié comme *terme-technique* avec une certitude de .4:

texte(s, «*recherche d'informations*»).
terme-technique(s)[.4].

Les alternatives sur le méta-contenu s'expriment comme nous l'avons déjà vu pour le contenu, par le biais du prédicat *alternative*. Par exemple, pour exprimer le fait que «*RI*» peut être *terme-technique* ou *abbréviation*, on aurait:

texte(s, «*RI*»).
alternative(s, s_1){.4}.
alternative(s, s_2){.6}.
terme-technique(s_1).
abbréviation(s_2).

Comme nous l'avons déjà mentionné à la section 3.1.2, les types ne sont pas nécessairement mutuellement exclusifs. Ainsi, pour l'exemple ci-dessus, il se pourrait fort bien que s soit en fait à la fois *terme-technique* et *abbréviation*. On aurait alors:

$texte(s, «RI»)$.
$terme\text{-}technique(s)$.
$abbréviation(s)$.

3.3.2 Logique

Les informations de l'ordre de la logique définissent les éléments qui forment la structure d'un document, c'est-à-dire son organisation en chapitres, sections, références, etc. Alors que les informations linguistiques sont généralement indépendantes du contexte, les informations logiques, quant à elles, perdent tout leur sens sans la structure dont elles font partie. Par exemple, si une abbréviation demeure une abbréviation même si elle est isolée du reste du texte, le titre d'un chapitre, par contre, n'a pas de sens sans le lien qui l'unit à ce chapitre.

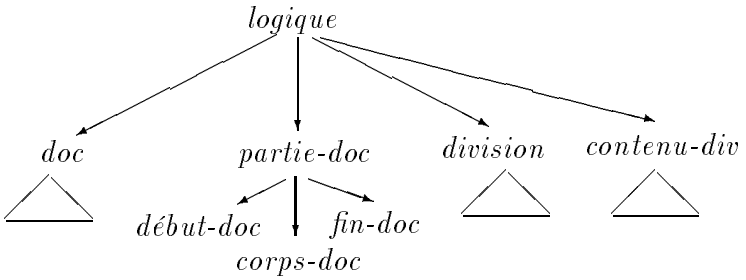


Figure 3.7. Types de contenu logique

La figure 3.7 résume les différents types de contenu logique, que nous explicitons ci-dessous. Il va sans dire qu'étant donné la grande diversité de documents, cette liste ne saurait être exhaustive. Nous nous concentrons surtout ici sur les textes de prose scientifique. Ces types font bien sûr l'objet de nombreuses contraintes traduisant des règles de composition; ces contraintes sont définies à la section 3.5.1.

a) Documents

La figure 3.8 résume les différents types de documents. Étant donné la grande diversité de ces types, il serait impossible d'en faire la liste exhaustive ici. D'ailleurs cette liste est

forcément ouverte, puisque les types peuvent toujours être raffinés, et que de nouveaux types de documents peuvent émerger, en particulier pour ce qui est des documents électroniques. Nous nous limitons donc ici aux documents explicatifs, et plus particulièrement aux communications scientifiques.

Nous distinguons les *documents* des *références* à des documents: tous deux peuvent avoir des attributs bibliographiques, tels qu'un auteur, une date de publication, etc., mais seul le *document* contient le texte du document. Les *réf-documents* ne sont que des références bibliographiques, auxquelles peuvent s'ajouter éventuellement un résumé.

Les *mono-documents* sont les ouvrages *monolithiques*, c'est-à-dire entièrement écrits par le(s) même(s) auteur(s). Il s'agit de: *article*, *livre*, *thèse*, et *rapport-technique*. Les *multi-documents* sont des documents composés d'autres documents dont les auteurs sont distincts. On retrouve sous cette classe: *revue* et *actes*.

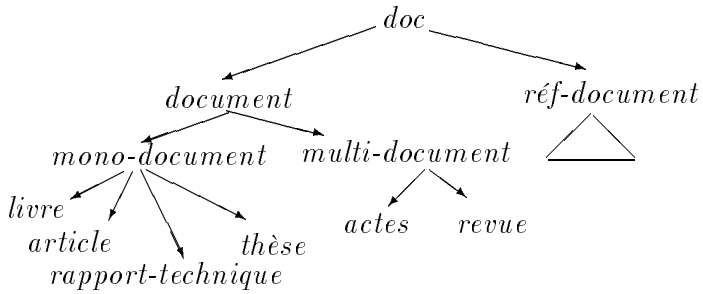


Figure 3.8. Types de documents

La hiérarchie sous *réf-document* est similaire à celle sous *document*: pour chaque sous-type de *document* il existe un sous-type correspondant sous *réf-document*, qui désigne cette fois une référence au document plutôt que le document lui-même.

b) Parties de documents

Les *documents* sont divisés en trois grandes parties, telles que montrées sous le type *partie-doc* à la figure 3.7. La partie *début-doc* regroupe les sections qui précèdent le corps du document, comme par exemple la préface ou la table des matières. La partie *corps-doc* constitue le corps du document. Son contenu varie grandement dépendamment du type de document: il peut s'agir de sections ou chapitres dans le cas de *mono-document*, ou même d'autres documents dans le cas de *multi-document*. Enfin, la partie *fin-doc* regroupe postface, bibliographie, annexes, etc.

c) Divisions

Chaque *partie-doc* contient une ou plusieurs *division* parmi celles montrées à la figure 3.9. Ces divisions comprennent les chapitres, sections, etc., qui apparaissent dans le document. Les divisions qui sont normalement créées automatiquement par les traitements de texte, comme par exemple la table des matières, l'index, etc., ne sont pas représentées ici. Les divisions de type *corps* ne peuvent apparaître qu'au sein d'un *corps-doc*, alors que les divisions de type *non-corps* apparaissent dans un *début-doc* ou un *fin-doc*.

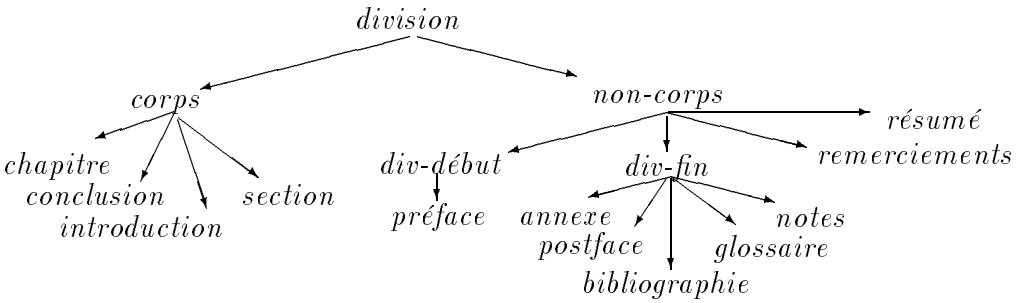


Figure 3.9. Types de divisions

Les types *chapitres* et *sections* apparaissent habituellement dans le corps du document. Les types *div-début* et *div-fin* concernent certains types de *divisions* qui ne peuvent apparaître qu'en début ou en fin de document, respectivement. Le *résumé* peut se retrouver à la fin ou au début du document. Il en va de même pour les *remerciements*, qui sont généralement placés à la fin pour les *articles* et au début pour les autres documents.

La relation *niveau* peut être utilisée pour donner la profondeur d'une division. Pour exprimer que la section 3.3.2 a une profondeur de 3, on aurait:

$section(s_{3.3.2})$.
$niveau(s_{3.3.2}, 3)$.

d) Contenu d'une division

Le contenu d'une *division* est donné par un ou plusieurs *contenu-div*, tels que montrés à la figure 3.10. Nous nous limitons encore une fois aux types pertinents pour la prose, et rejetons donc des éléments comme les strophes, vers, etc.

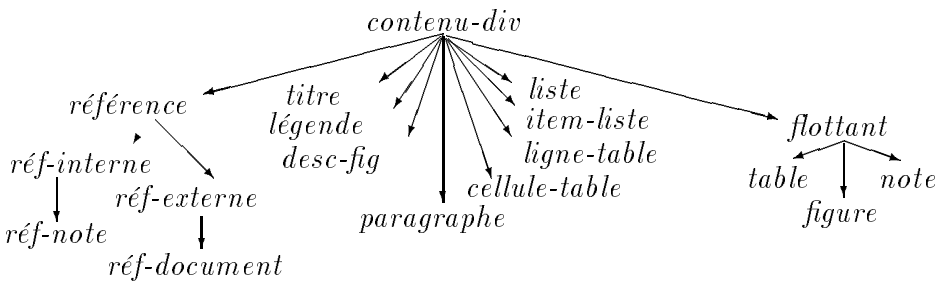


Figure 3.10. Hiérarchie du contenu d'une division

– Les éléments de type *flottant* n'apparaissent pas nécessairement dans le document imprimé à la même position par rapport aux autres éléments, que la position où ils sont dans le document électronique. Ainsi, pour être placée en haut d'une page, une figure peut être avancée de quelques paragraphes voire même couper un paragraphe en deux. Les éléments d'information de ce type sont: les *note*s, les *figure*s et les *table*s.

– Les *note*s sont des annotations au texte qui apparaissent soit en bas de page, soit dans une division *notes* à la fin du document.

– Une *figure* indique la position d'un graphique, d'une illustration ou d'une figure. Les *figure*s sont des éléments logiques qui contiennent à leur tour d'autres éléments: en particulier, en plus d'une image, elles peuvent contenir un *titre*, une *légende* ou une *desc-fig*. Un exemple complet de *figure* est donné à la section 3.5.1.

– une *table* contient du texte affiché sous une forme tabulaire, en lignes et colonnes. Le contenu d'une *table* est organisé en lignes (*ligne-table*), chacune contenant une ou plusieurs cellules (*cellule-table*). Les *table*s peuvent aussi contenir un *titre* et une *légende*.

  Voici ci-dessous un exemple de déclaration de *table* comportant deux lignes et deux colonnes. Un exemple plus complet est donné à la section 3.5.1.

  $table(s).$
  $lignes(s, 2).$
  $colonnes(s, 2).$

– on distingue deux types de *référence*s: les références à une autre partie du même document (*réf-interne*) et les références à d'autres documents (*réf-externe*). Le type *réf-document* est un type particulier de *réf-externe*, où l'objet référencé est un do-

cument entier; quant à *réf-note*, il s'agit d'un renvoi à une *note*. Des exemples sont donnés plus loin à la section 3.4.2.

- le type *liste* sert à grouper une liste de *item-liste* ensemble. À noter que dans le document originel, les items peuvent être clairement séparés du texte principal ou non.

- *paragraphe*. Un paragraphe de texte.

- *titre*. Le titre d'une *division*, d'un *document*, d'une *figure* ou d'une *table*.

e) Incertitude logique

Tout comme pour les informations de type *linguistique*, chaque information de type *logique* peut avoir une certitude associée ou faire l'objet de multiples alternatives. Ainsi, dans l'exemple ci-dessous, s a deux types possibles, *titre* ou *paragraphe*.

$texte(s, \text{«}Les\ carnets\ rouges\text{»})$.
$alternative(s, s_1)\{.4\}$.
$alternative(s, s_2)\{.6\}$.
$titre(s_1)$.
$paragraphe(s_2)$.

3.3.3 Attributs

Nous avons déjà vu de façon informelle dans les sections précédentes plusieurs des attributs de $\mathcal{L}$. Nous reprenons et complétons la liste dans cette section. Notons que bon nombre de ces attributs ont été inspirés de *Bibliographic-Data*, une ontologie intégrée dans *Ontolingua* [Gru93, Gru92].

a) Types d'attributs

La figure 3.11 présente certains types d'information particuliers qui peuvent servir de valeur à une relation d'attribut. Ce sont:

- *agent*. Une entité qui a une certaine relation avec le passage de texte ou le document. Il peut s'agir par exemple de l'auteur d'un document ou d'une citation. Une *personne* est un individu, tandis qu'un *organisme* peut être par exemple une maison d'édition. La propriété *nom-agent* permet d'associer un nom à un *agent*. Par exemple:

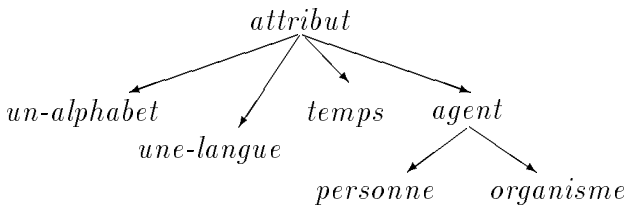


Figure 3.11. Types d'attributs

personne(*jfc*).
nom-agent(*jfc*, «*James Fenimore Cooper*»).

– *temps*. Un point dans le temps, comme par exemple la date de publication d'un document. La propriété *valeur-temps* permet d'y associer une chaîne de caractères.

– *un-alphabet*. L'alphabet dans lequel le texte est encodé. Les constantes suivantes sont définies:

un-alphabet(*Romain*).
un-alphabet(*Cyrillique*).

– *une-langue*. La langue du texte. Les constantes suivantes sont définies:

un-alphabet(*anglais*).
un-alphabet(*français*).

Les types *agent* et *temps* sont surtout utiles lorsqu'il s'agit de comparer plusieurs documents entre eux; par exemple, grâce à l'existence d'une entité *auteur* pour chaque document, il est possible de lier entre eux tous les documents écrits par un même auteur. De plus, de cette façon, des ouvrages écrits sous des noms de plume différents ou avec un pseudonyme peuvent être liés au même auteur. Lorsque l'on n'est intéressé que par le nom d'un agent, par contre, cette indirection peut s'avérer gênante. C'est pourquoi tous les attributs qui utilisent l'un ou l'autre de ces types permettent aussi d'utiliser directement une chaîne de caractères (voir section suivante).

b) Attributs bibliographiques

Les attributs bibliographiques peuvent être associés à un *doc*, que ce soit un document ou une référence à un document. De nombreux formalismes ou ontologies existent déjà pour l'expression de références bibliographiques. Nous nous inspirons ici de l'ontologie *Bibliographic-Data*, déjà citée, de BibTeX, le standard LaTeX, ainsi que de TEI [TEI94]. Voici les différents attributs bibliographiques:

- *auteur*, le ou les auteurs du document;
- *titre-doc*, le titre du document;
- *date*, la date de publication;
- *publié-par*, l'*organisme* responsable de la publication.;
- *publié-à*, l'endroit où la publication a eu lieu;
- *édité-par*, l'éditeur;
- *sélection*, précise davantage la portée de la référence bibliographique, qu'il s'agisse d'une sélection de pages, du volume ou du numéro d'un périodique, etc.;
- *id*, un numéro standard pour le document, par exemple, un numéro ISBN.

Ces attributs ont les contraintes de types suivantes:

$$auteur(\sigma,\alpha) \supset (doc(\sigma) \wedge (chaîne(\alpha) \vee agent(\alpha))). \quad (3.22)$$
$$titre\text{-}doc(\sigma,\alpha) \supset (doc(\sigma) \wedge (chaîne(\alpha) \vee titre(\alpha))). \quad (3.23)$$
$$date(\sigma,\alpha) \supset (doc(\sigma) \wedge (chaîne(\alpha) \vee temps(\alpha))). \quad (3.24)$$
$$publié\text{-}par(\sigma,\alpha) \supset (doc(\sigma) \wedge (chaîne(\alpha) \vee organisme(\alpha))). \quad (3.25)$$
$$publié\text{-}à(\sigma,\alpha) \supset (doc(\sigma) \wedge chaîne(\alpha)). \quad (3.26)$$
$$édité\text{-}par(\sigma,\alpha) \supset (doc(\sigma) \wedge (chaîne(\alpha) \vee agent(\alpha))). \quad (3.27)$$
$$sélection(\sigma,\alpha) \supset (doc(\sigma) \wedge chaîne(\alpha)). \quad (3.28)$$
$$id(\sigma,\alpha) \supset (doc(\sigma) \wedge chaîne(\alpha)). \quad (3.29)$$

On voit que les attributs bibliographiques peuvent tous accepter comme valeur une chaîne de caractères. Dans l'exemple ci-dessous, chaque attribut du document d est représenté par une chaîne de caractères.

réf-document(d).
auteur$(d,$ «*James Fenimore Cooper*»$)$.
titre-doc$(d,$ «*Le dernier des Mohicans*»$)$.
date$(d,$ «*1826*»$)$.

Ces attributs pourraient aussi être représentés à l'aide des types d'attribut définis à la section précédente. Sachant que *jfc*, représente «*James Fenimore Cooper*» tel que défini ci-dessus, on peut réécrire la relation *auteur* comme suit:

$$auteur(d, jfc).$$

Certaines informations peuvent apparaître aussi bien en tant qu'attributs bibliographiques qu'en tant que contenu textuel du document. C'est le cas notamment des informations qui se retrouvent sur une page titre, comme le titre, auteur(s), etc. On peut exprimer le fait qu'un attribut bibliographique coïncide avec un élément du contenu textuel, en remplaçant la chaîne de caractères par un objet de type *contenu-textuel*. Par exemple, sachant que *dm* est un élément de la page titre de *d* donnant le titre du document, on peut réécrire l'attribut *titre-doc* de *d* comme suit:

$$titre(dm).$$
$$texte(dm, \text{«}Le\ dernier\ des\ Mohicans\text{»}).$$
$$titre\text{-}doc(d, dm).$$

Notons cependant que les informations apparaissant sur la page titre peuvent être distinctes des attributs bibliographiques. Ces derniers sont vus comme étant plus «*fiables*», puisqu'ils requièrent généralement une intervention humaine pour être ainsi identifiés. Toutefois, dans le cas où l'on ne dispose pas des attributs bibliographiques, les informations de contenu textuel peuvent être utilisées pour déduire titre, auteur, etc.

c) Autres attributs

Nous avons vu tout au long de la section 3.3.1 plusieurs exemples d'attributs portant sur les informations de type *linguistique*. Ce sont:

- *auteur-citation*, qui est responsable d'une *citation*;
- *abbr* et *expan*, respectivement l'abbréviation ou l'expansion d'une expression;
- *langue*, *alphabet*, respectivement la langue ou l'alphabet d'un texte;
- *externe*, un lien vers un fichier externe, qui contient la repésentation d'une image, vidéo, etc., mais aussi éventuellement d'un texte.

Les attributs suivants s'appliquent aux informations de type *logique*:

- *niveau*, indique le niveau hiérarchique d'une *division*;

– *lignes* et *colonnes*, indiquent respectivement le nombre de lignes et de colonnes dans une *table*.

Enfin, deux attributs sont définis pour les *attributs*: il s'agit de *nom-agent* et de *valeur-temps*, qui associent tous deux une chaîne de caractères à un *agent* ou à un *temps*, respectivement.

Ces attributs ont les contraintes de types sur leurs arguments telles que définies par les règles 3.30 à 3.40.

$$\textit{auteur-citation}(\sigma,\alpha) \supset (\textit{contenu-textuel}(\sigma) \wedge (\textit{chaîne}(\alpha) \vee \textit{agent}(\alpha))). \quad (3.30)$$
$$\textit{expan}(\sigma_1,\sigma_2) \supset (\textit{abbréviation}(\sigma_1) \wedge \textit{expansion}(\sigma_2)). \quad (3.31)$$
$$\textit{abbr}(\sigma_2,\sigma_1) \supset (\textit{expansion}(\sigma_2) \wedge \textit{abbréviation}(\sigma_1)). \quad (3.32)$$
$$\textit{langue}(\sigma,\alpha) \supset (\textit{contenu-textuel}(\sigma) \wedge \textit{une-langue}(\alpha)). \quad (3.33)$$
$$\textit{alphabet}(\sigma,\alpha) \supset (\textit{contenu-textuel}(\sigma) \wedge \textit{un-alphabet}(\alpha)). \quad (3.34)$$
$$\textit{externe}(\sigma,\tau) \supset (\textit{contenu-signifiant}(\sigma) \wedge \textit{chaîne}(\tau)). \quad (3.35)$$
$$\textit{niveau}(\sigma,\alpha) \supset (\textit{division}(\sigma) \wedge \textit{nombre}(\alpha)). \quad (3.36)$$
$$\textit{lignes}(\sigma,\alpha) \supset (\textit{table}(\sigma) \wedge \textit{nombre}(\alpha)). \quad (3.37)$$
$$\textit{colonnes}(\sigma,\alpha) \supset (\textit{table}(\sigma) \wedge \textit{nombre}(\alpha)). \quad (3.38)$$
$$\textit{nom-agent}(\alpha,\tau) \supset (\textit{agent}(\alpha) \wedge \textit{chaîne}(\tau)). \quad (3.39)$$
$$\textit{valeur-temps}(\alpha,\tau) \supset (\textit{temps}(\alpha) \wedge \textit{chaîne}(\tau)). \quad (3.40)$$

Le prédicat *nombre* est similaire à *chaîne*; il indique si un élément est numérique.

d) Certitude sur les attributs

Dans l'exemple ci-dessous, on a une certitude de .8 que l'auteur de la *citation* «*L'état c'est moi*» est «*Louis XIV*».

$\textit{citation}(s).$
$\textit{texte}(s, \text{«}\textit{L'état c'est moi}\text{»}).$
$\textit{auteur-citation}(s, \text{«}\textit{LouisXIV}\text{»})[.8].$

3.3.4 Non-textuel

Nous avons déjà vu en quoi pouvait consister les informations de contenu *non-textuel*: il s'agit ici de permettre de préciser davantage leur type. Certains des types d'images et d'autres média sont montrés à la figure 3.12; cette liste est bien entendu loin d'être exhaustive.

Voici un exemple d'information non-textuelle, où une image de type *bitmap* est stockée dans le fichier «*fig1.bmp*»:

image-bitmap(*fig1*).
externe(*fig1*, «*fig1.bmp*»).

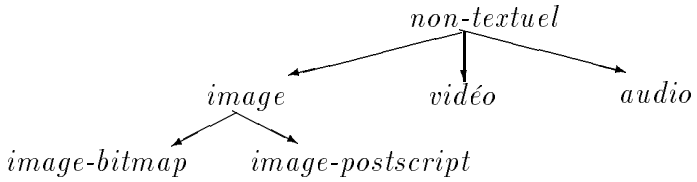


Figure 3.12. Types de contenu non-textuel

Il peut arriver qu'une même image soit codée sous plusieurs formats; on représente alors ces différents formats à l'aide d'*alternatives*.

3.3.5 Méta-sémantique

La figure 3.13 montre quelques types d'information *méta-sémantique*.

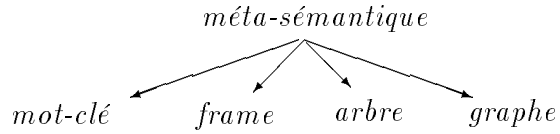


Figure 3.13. Types de contenu sémantique

Tout comme pour les informations non-textuelles, il est possible qu'une même représentation sémantique existe sous plusieurs formats. Notre approche est la même que ci-dessus pour les types non-textuels: elle consiste à définir chaque représentation par une *alternative*. Ainsi, dans l'exemple ci-dessous, la représentation sémantique de «*Le Renard Subtil*» est donnée par mots-clés (rs_1) et par graphe (rs_2):

$contenu\text{-}textuel(s)$.
$texte(s, \text{«}Le\ Renard\ Subtil\text{»})$.
$sém(s, rs)$.

$alternative(rs, rs_1)$.
$alternative(rs, rs_2)$.
$mot\text{-}clé(rs_1)$.
$graphe(rs_2)$.

3.3.6 Thèmes

Il est intéressant de constater que plusieurs linguistes sont arrivés à la conclusion – vérifiée expérimentalement en recherche d'informations – que des segments de texte plus petits que le texte lui-même peuvent aussi avoir des thèmes propres; et que de plus, les textes n'ont pas nécessairement un thème unique [HS84]. Cette idée est reprise dans notre modèle, puisque chaque thème est associé à une portion de texte, et qu'une portion de texte peut comporter plusieurs thèmes.

Concrètement, les thèmes dans $\mathcal{L}$ sont des informations de type *contenu-sémantique* – le plus souvent des mots-clés, mais il peut aussi s'agir de représentations sémantiques plus poussées comme des graphes ou des arborescences – qui sont liées à un élément de type *contenu-signifiant* par le biais de la relation *thème*.

$$thème(\sigma, \varsigma) \quad \supset \quad (contenu\text{-}signifiant(\sigma) \wedge contenu\text{-}sémantique(\varsigma)). \tag{3.41}$$

À cette relation peut s'ajouter un poids qui indique l'importance relative d'un thème pour sa section. Ce poids peut être exprimé par le traditionnel calcul *tf.idf* ou par d'autres méthodes. Ainsi dans l'exemple ci-dessous, *mohican* et *guerre* sont tous deux thèmes de d, mais *mohican* est considéré comme «*meilleur*» ou plus significatif puisque $.4 \leq .8$.

$thème(d, mohican)\{.8\}$.
$thème(d, guerre)\{.4\}$.

L'exemple ci-dessus montre comment exprimer la conjonction entre thèmes – puisqu'ici les deux thèmes s'appliquent en même temps à d. Cette conjonction n'est pas dans le sens des thèmes «*composés*», comme par exemple *recherche* et *information* qui forment ensemble «*recherche d'informations*». Les thèmes composés doivent s'exprimer par des représentations sémantiques appropriées.

Il est aussi souhaitable dans certains cas d'exprimer la disjonction entre thèmes. Si cette notion de disjonction semble naturelle dans un modèle booléen, elle n'est pas très intuitive du point de vue d'un utilisateur, puisqu'elle lui cache souvent les vraies raisons de ces alternatives. Comme la disjonction représente généralement une incertitude, nous privilégions plutôt de la représenter par des alternatives sur le type d'information touché par l'incertitude. Par exemple, pour exprimer que le thème correspondant à «*tribu des Cinq Nations*» est soit *mohican*, soit *mohawk*, on aura:

contenu-textuel(s).
texte$(s,$ « *tribu des Cinq Nations* »$)$.
thème(s,t).
alternative$(t, mohican)$.
alternative$(t, mohawk)$.

Cette représentation signifie qu'il y a ambiguïté sémantique sur t, c'est-à-dire que la même chaîne de caractères a deux interprétations sémantiques différentes. Une ambiguïté sur la chaîne de caractères elle-même et les thèmes qui en découlent s'exprimerait comme suit:

contenu-textuel(s).
alternative(s, s_1).
alternative(s, s_2).
texte$(s_1,$ « *Mohicans* »$)$.
texte$(s_2,$ « *Mohawks* »$)$.
thème$(s_1, mohican)$.
thème$(s_2, mohawk)$.

3.3.7 Connaissances

Nous considérons comme *connaissance* toute information sémantique qui est externe au document, c'est-à-dire qui ne découle pas directement de sa représentation sémantique, mais qui est plutôt du ressort de la modélisation du domaine.

On peut classer les relations entre deux concepts en deux grands groupes : les relations *hiérarchiques* et *non-hiérarchiques* [SDV95, pp15–19]. Les relations hiérarchiques incluent l'*hyponymie*, qui permet la définition de taxonomies, et la *méronymie*, qui exprime le lien de composition. Les relations non-hiérarchiques comprennent essentiellement les synonymes et antonymes. Dans $\mathcal{L}$, seules deux relations sont considérées: la relation *implique* (hyponymie), et la relation *équivalent* (synonymie). Rien n'empêche toutefois la définition de relations plus complexes par le biais de règles.

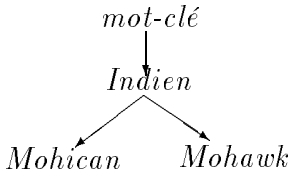
On exprime le fait qu'un élément en implique un autre par la relation *implique*; ceci permet notamment de définir une hiérarchie de *généralisation/spécialisation* des concepts. Ainsi, en supposant une représentation par *mot-clés*, la figure 3.14 définit quelques types de contenu sémantique possibles. On voit à gauche la représentation arborescente et à droite l'équivalent dans $\mathcal{L}$.

La relation *implique* est réflexive (règle 3.42) et transitive (règle 3.43). Les poids obtenus par transitivité sont forcément inférieurs ou égaux aux poids pris séparément, ce qui

s'exprime par $\delta_3 \leq \delta_1$ et $\delta_3 \leq \delta_2$ dans la règle 3.43.

$$implique(\varsigma, \varsigma)\{\delta_\tau\}. \tag{3.42}$$

$$\begin{aligned}(&implique(\varsigma_1, \varsigma_2)\{\delta_1\} \wedge \\ &implique(\varsigma_2, \varsigma_3)\{\delta_2\}) \supset implique(\varsigma_1, \varsigma_3)\{\delta_3\} \wedge (\delta_3 \leq \delta_1) \wedge (\delta_3 \leq \delta_2).\end{aligned} \tag{3.43}$$



$implique(Mohican, Indien).$
$implique(Mohawk, Indien).$

Figure 3.14. Exemple de connaissances

L'*équivalence* mesure le degré de synonymie entre deux éléments. En plus d'être réflexive et transitive, elle est aussi symétrique.

$$équivalent(\varsigma, \varsigma)\{\delta_\tau\}. \tag{3.44}$$

$$\begin{aligned}(&équivalent(\varsigma_1, \varsigma_2)\{\delta_1\} \wedge \\ &équivalent(\varsigma_2, \varsigma_3)\{\delta_2\}) \supset équivalent(\varsigma_1, \varsigma_3)\{\delta_3\} \wedge \delta_3 \leq \delta_1 \wedge \delta_3 \leq \delta_2.\end{aligned} \tag{3.45}$$

$$équivalent(\varsigma_1, \varsigma_2)\{\delta\} \supset équivalent(\varsigma_2, \varsigma_1)\{\delta\}. \tag{3.46}$$

3.4 Représentation de la structure

Nous avons vu jusqu'à présent quels étaient les types d'information dans les documents et comment ils se traduisaient dans $\mathcal{L}$. Nous voyons maintenant comment ces informations peuvent s'organiser entre elles.

Le contenu d'un document est donné par un ensemble de composants qui sont à leur tour définis en donnant leurs composants. La structure de base d'un document est donc donnée par une relation de *composition*, à laquelle s'ajoute une relation d'*ordonnancement* qui donne l'ordre – pas forcément linéaire – dans lequel le document doit être parcouru.

Nous distinguons trois grands types de structure: la structure *linguistique*, qui donne l'organisation des informations de *contenu-textuel* et *linguistique*, la structure *logique* qui donne l'organisation des informations *logiques*, et enfin, la structure *de discours*, qui donne quant à elle l'organisation des *idées*. Ces structures sont décrites en détails dans les sections qui suivent. On remarquera que la plupart des règles définies ci-dessous comportent la certitude maximale sur leurs prédicats; ceci implique que les propriétés des relations de composition et d'ordonnancement ne sont pas connues quand elles sont incertaines.

3.4.1 Structure linguistique

Nous avons déjà mentionné à la section 3.2.1 comment la combinaison d'informations de *contenu-textuel* était essentielle à la bonne compréhension d'un texte. Nous considérons ici principalement deux types de relations structurelles dites «*linguistiques*»: la relation de *séquentialité*, qui donne l'ordre dans lequel doit être compris le texte, et la relation de *composition*, qui permet de grouper les éléments textuels. Nous introduisons de plus une relation de dépendance, qui est inspirée des théories de grammaires de dépendances [Gen91].

a) Composition linguistique

Les informations de contenu textuel peuvent être *composées*, c'est-à-dire formées à partir d'autres informations. Cette notion correspond intuitivement au groupement de morphèmes pour former des mots, au groupement de mots pour former des syntagmes, qui forment à leur tour des phrases, etc.

La relation de composition est exprimée par *part*. Dans l'exemple ci-dessous, «*recherche d'informations*» (s) est composé de «*recherche*», «*d'*» et «*informations*».

$part(s, s_1)$. $texte(s_1, \text{«}recherche\text{»})$.
$part(s, s_2)$. $texte(s_2, \text{«}d'\text{»})$.
$part(s, s_3)$. $texte(s_3, \text{«}informations\text{»})$.

Puisque les documents peuvent aussi contenir des informations non-textuelles, la relation *part* est étendue à toute information de type *contenu-signifiant*. Toutefois, si une image ou un symbole graphique peuvent parfaitement apparaître au sein d'un paragraphe, les informations de type *linguistique*, elles, sont restreintes à des composants qui sont eux aussi de type *linguistique*. Ces contraintes sont exprimées par les deux règles suivantes:

$$part(\sigma_1, \sigma_2)[\mu] \supset (contenu\text{-}signifiant(\sigma_1) \wedge contenu\text{-}signifiant(\sigma_2)). \quad (3.47)$$

$$(part(\sigma_1, \sigma_2)[\mu] \wedge linguistique(\sigma_1)) \supset linguistique(\sigma_2). \quad (3.48)$$

La relation *part* possède en outre les propriétés suivantes: elle est *antisymmétrique* (règle 3.49), *irréflexive* (règle 3.50), et *injective* (règle 3.51).

$$antisymétrique(part). \quad (3.49)$$

$$irréflexive(part). \quad (3.50)$$

$$injective(part). \quad (3.51)$$

On aura remarqué que la relation *part* n'est pas transitive; ceci permet principalement de différencier les descendants *directs* de ceux qui sont plus bas dans la hiérarchie des composants. À cet effet, la relation *part-trans*, qui elle est *transitive* (règle 3.53), peut être déduite à partir de *part* (règle 3.52). Cette relation est de plus *antisymmétrique* (règle 3.54) et *irréflexive* (règle 3.54), mais pas *injective*.

$$part(\sigma_1, \sigma_2) \supset part\text{-}trans(\sigma_1, \sigma_2). \tag{3.52}$$
$$transitive(part\text{-}trans). \tag{3.53}$$
$$antisymétrique(part\text{-}trans). \tag{3.54}$$
$$irréflexive(part\text{-}trans). \tag{3.55}$$

En pratique, seule la relation *part* est utilisée dans la déclaration des faits, alors que la relation *part-trans* est utilisée dans les règles pour retrouver tous les composants d'une structure, qu'ils soient descendants directs on non.

b) Ordonnancement linguistique

L'ordonnancement linguistique est dit «*séquentiel*», c'est-à-dire que dans la lecture d'un texte, un élément doit *suivre* un autre. Si cette caractéristique est commune à toutes les langues, cela ne signifie pas que cet ordre soit le même pour toutes ces langues; ainsi, dans les langues romanes on lit de gauche à droite, alors qu'avec d'autres langues, comme l'arabe, le lecture se fait de droite à gauche.

La relation *séq* est utilisée pour indiquer qu'un élément suit immédiatement un autre dans le texte. Ainsi, pour ajouter la séquence linguistique à la représentation de «*recherche d'informations*» ci-dessus, on aurait:

$$séq(s_1, s_2).$$
$$séq(s_2, s_3).$$

La relation *séq* lie deux informations de type *contenu-signifiant*, puisque le texte peut très bien être entremêlé d'éléments graphiques ou autres informations non-textuelles:

$$séq(\sigma_1, \sigma_2)[\mu] \supset (contenu\text{-}signifiant(\sigma_1) \wedge contenu\text{-}signifiant(\sigma_2)). \tag{3.56}$$

La relation *séq* possède les mêmes contraintes que *part*, c'est-à-dire qu'elle est *antisymmétrique* (règle 3.57), *irréflexive* (règle 3.58), et *injective* (règle 3.59). De plus, chaque élément d'information peut avoir au plus un successeur avec *séq* (règle 3.60).

$$antisymétrique(séq). \tag{3.57}$$
$$irréflexive(séq). \tag{3.58}$$

$$injective(séq). \tag{3.59}$$

$$(séq(\sigma_1, \sigma_2) \wedge séq(\sigma_1, \sigma_3)) \supset \sigma_2 = \sigma_3. \tag{3.60}$$

Tout comme pour la relation de composition, il existe une relation de séquence *transitive*, appelée *séq-trans*, qui est définie à partir de *séq*, et qui est également *antisymmétrique* et *irréflexive*:

$$séq(\sigma_1, \sigma_2) \supset séq\text{-}trans(\sigma_1, \sigma_2). \tag{3.61}$$

$$transitive(séq\text{-}trans). \tag{3.62}$$

$$antisymétrique(séq\text{-}trans). \tag{3.63}$$

$$irréflexive(séq\text{-}trans). \tag{3.64}$$

Le sens de *séq-trans* est à prendre comme une relation de *précédence*, c'est-à-dire qu'elle permet de dire lequel de deux éléments appararaît avant l'autre dans le document.[9]

On définit enfin le prédicat *adjacent*, qui indique que deux éléments sont voisins mais sans préciser lequel précède l'autre dans le texte. Cette relation est évidemment de peu d'utilité dans la description du document, puisque les éléments sont toujours effectivement en séquence, mais est utile pour la définition des règles d'indexation.

$$séq(\sigma_1, \sigma_2) \supset adjacent(\sigma_1, \sigma_2). \tag{3.65}$$

$$symétrique(adjacent). \tag{3.66}$$

$$transitive(adjacent). \tag{3.67}$$

c) Dépendance linguistique

La relation *dépendant* permet d'exprimer le fait qu'une information est *subordonnée* à une autre. On l'utilisera le plus souvent pour refléter l'arbre d'analyse d'une expression, en identifiant les relations de dépendances entre les items grammaticaux, e.g. entre un déterminant et un nom. Par exemple, toujours pour la représentation de «*recherche d'informations*»; sachant que s_2 correspond à «*d'*» et s_3 à «*informations*», on peut exprimer le fait que s_2 est dépendant de s_3 par:

$$dépendant(s_2, s_3).$$

La relation *dépendant* n'a de sens que sur des éléments de *contenu-textuel*:

$$dépendant(\sigma_1, \sigma_2)\{\delta\} \supset (contenu\text{-}textuel(\sigma_1) \wedge contenu\text{-}textuel(\sigma_2)). \tag{3.68}$$

9. Ceci est vrai pour toute information de *contenu-signifiant*, à l'exception des *flottant*s (voir section suivante sur l'ordonnancement *logique*).

Le poids δ représente la «*force*» de la relation de dépendance entre les deux éléments. La relation *dépendant* est *antisymmétrique* (règle 3.69) et *irréflexive* (règle 3.70).

$$antisymétrique(dépendant). \tag{3.69}$$
$$irréflexive(dépendant). \tag{3.70}$$

On peut rarement trouver cette information dans les textes initiaux, aussi la relation *dépendant* est-elle le plus souvent déduite, en utilisant les règles définies dans le chapitre suivant.

d) Ambiguïtés sur la structure linguistique

Les relations de *composition* et d'*ordonnancement* peuvent être accompagnées d'une mesure de *certitude*. En reprenant notre exemple de «*recherche d'informations*», où s_1 représente «*recherche*» et s_3 «*informations*», on peut exprimer le fait que s_1 est suivi de s_3 avec une certitude de .8 par:

$$séq\text{-}trans(s_1, s_3)[.8].$$

On peut s'interroger sur le sens d'une certitude portant sur les relations *part* et *séq*. Ce sens n'est généralement pas très clair, sauf peut-être dans le cas d'un document obtenu par reconnaissance de caractères, où la juxtaposition des mots n'est pas toujours reconnue, et où des informations erronées – comme par exemple une tache d'encre – peuvent s'ajouter à la représentation. Une incertitude sur ces deux relations a surtout le désavantage de limiter l'inférence de nouveaux faits sur le document, puisque les règles que nous avons données ci-dessus ne s'appliquent qu'avec des certitudes maximum. Lorsque c'est possible, il est donc préférable de donner toutes les *alternatives* à une représentation, avec éventuellement les *poids* qui les accompagnent.

Ainsi, soit l'expression «*Il regarde la fille avec le télescope*», où s_1 représente «*Il regarde*», s_2, «*la fille*» et s_3, «*avec le télescope*». Dans la représentation ci-dessous, on pose a, comme étant un objet représentant les deux analyses possibles de cette phrase. La première alternative, «s_1 (s_2 s_3)», est donnée par a_1, et la seconde, «(s_1 s_2) s_3», par a_2:

$texte(s_1, \text{«}Il\ regarde\text{»})$.
$texte(s_2, \text{«}la\ fille\text{»})$.
$texte(s_3, \text{«}avec\ le\ télescope\text{»})$.
$séq(s_1, s_2)$.
$séq(s_2, s_3)$.

$alternative(a, a_1)$.
$alternative(a, a_2)$.
$part(a_1, s_1)$.
$part(a_1, s_{2+3})$.
$part(s_{2+3}, s_2)$.
$part(s_{2+3}, s_3)$.
$part(a_2, s_{1+2})$.
$part(a_2, s_3)$.
$part(s_{1+2}, s_1)$.
$part(s_{1+2}, s_2)$.

3.4.2 Structure logique

La structure de composition et d'ordonnancement logique est essentiellement une extension à la structure linguistique, c'est-à-dire qu'elle partage les mêmes relations, *part* et *séq*, sauf que ces relations s'appliquent non plus à des mots, syntagmes, etc., mais à des sections, chapitres, etc.

a) Composition logique

De la même façon que les informations linguistiques peuvent se combiner entre elles, les informations logiques sont elles aussi formées à partir d'autres informations. Les composantes sont de type *contenu-signifiant*, c'est-à-dire qu'elles peuvent être de nature logique, linguistique ou même non-textuelle (voir règle 3.47). Des contraintes supplémentaires quant à la composition logique sont données à la section 3.5.1.

Dans l'exemple ci-dessous, la *section* $s_{3.4.2}$ est formée d'un *titre* (s_t) et d'un *paragraphe* (s_p).

$section(s_{3.4.2})$.
$part(s_{3.4.2}, s_t)$.
$part(s_{3.4.2}, s_p)$.
$titre(s_t)$.
$paragraphe(s_p)$.
$texte(s_t, \text{«}Structure\ logique\text{»})$.
$texte(s_p, \text{«}La\ structure\ de\ composition...\text{»})$.

Le prédicat *contient-texte* indique si un élément de contenu textuel σ, ou l'une de ses

parties σ_s, contiennent la chaîne τ. Cette contrainte est vérifiée par les règles suivantes:

$$texte(\sigma, \tau) \supset contient\text{-}texte(\sigma, \tau). \quad (3.71)$$

$$(part\text{-}trans(\sigma, \sigma_s) \wedge texte(\sigma_s, \tau)) \supset contient\text{-}texte(\sigma, \tau). \quad (3.72)$$

b) Ordonnancement logique

La séquence logique, tout comme la séquence linguistique, s'exprime par le prédicat *séq*, et permet de fixer l'ordre des *paragraphe*s, *division*s, etc. Les titres de sections, puisqu'ils s'insèrent à un endroit précis du discours, sont aussi en relation de séquence avec le reste du texte. Le titre de cette section, s_t est ainsi en séquence avec le premier paragraphe, s_p:

$$séq(s_t, s_p).$$

En général il est préférable d'exprimer la séquence au plus haut niveau possible dans la hiérarchie de composition. Ainsi, même si le titre de la section $s_{3.4.3}$ suit le dernier paragraphe de $s_{3.4.2}$, la relation *séq* est plutôt exprimée entre ces deux sections.

$$séq(s_{3.4.2}, s_{3.4.3}).$$

La relation *séq-trans* peut être déduite pour les composants à partir de leurs parents, ou vice-versa pour les parents à partir de leurs composants. Toutes les informations qui sont en séquence avec le parent le sont aussi le fils (règle 3.73). De la même façon, toutes les informations qui sont en séquence avec un fils – pour autant qu'elles ne soient pas elles-mêmes filles du parent – sont aussi en séquence avec le parent (règle 3.74).

$$part\text{-}trans(\sigma_1, \sigma_2) \wedge séq\text{-}trans(\sigma_1, \sigma_3) \wedge \neg part\text{-}trans(\sigma_2, \sigma_3) \supset séq\text{-}trans(\sigma_1, \sigma_3). \quad (3.73)$$

$$part\text{-}trans(\sigma_1, \sigma_2) \wedge séq\text{-}trans(\sigma_2, \sigma_3) \wedge \neg part\text{-}trans(\sigma_1, \sigma_3) \supset séq\text{-}trans(\sigma_1, \sigma_3). \quad (3.74)$$

Le type *flottant*, qui comprend rappelons-le les *figure*s, *table*s et *note*s est le seul qui ne peut être lié par une relation *séq* ou *séq-trans* (règle 3.75).

$$(séq(\sigma_1, \sigma_2) \vee séq\text{-}trans(\sigma_1, \sigma_2)) \supset (\neg flottant(\sigma_1) \wedge \neg flottant(\sigma_2)).$$

En plus de la *séquence*, on peut exprimer au niveau logique la *référence*, c'est-à-dire le renvoi à d'autres parties du même texte ou d'un autre document. Il existe plusieurs motivations pour utiliser une référence dans un document [MF96]:

– pour citer un passage ou une citation («*il est dit dans ...* »);

- pour établir une association avec un autre document (« *notre approche est similaire à ...* »);
- pour une description analytique (« *comme le montre la figure ...* »);
- un renvoi à une définition (« *la recherche d'informations (cf section ...)* »).

Malheureusement les langages de représentation de documents actuels ne permettent pas de distinguer ces divers usages, et malgré tout l'intérêt que cela pourrait comporter, une telle analyse pourrait s'avérer très complexe. Nous nous limitons donc à un seul lien de référence. Une information plus précise peut être déduite en examinant le type des éléments en présence: ainsi les différentes références énumérées ci-dessus peuvent être retrouvées par les types *citation*, *réf-document*, *figure*, *glose*, etc. De plus, comme nous le verrons au chapitre suivant, l'analyse des expressions utilisées dans les références peut fournir des indices supplémentaires.

Les références ont toujours une *ancre* dans le texte, qui est donnée par un élément logique de type *référence*, et qui peut contenir du texte qui décrit l'objet référencé. Il existe deux types de *référence*: les *réf-interne*, dont la portée est à l'intérieur du même document, et les *réf-externe*, qui pointent sur des éléments dans d'autres documents.

Les *référence* sont liées à l'élément logique auxquel elles réfèrent par la relation *réf*. Dans l'exemple ci-dessous, s_2 est une référence à $s_{3.4.3}$:

$texte(s_1, \text{« voir la »})$.
$réf\text{-}interne(s_2)$.
$texte(s_2, \text{« section suivante »})$.
$réf(s_2, s_{3.4.3})$.

La relation *réf* n'accepte comme premier argument que des *référence*, et comme second, que des informations de type *logique* (règle 3.75). Les *réf-interne*s ne peuvent référencer que des éléments compris dans une même unité structurelle (règle 3.76), alors que pour les *réf-externe*s, c'est le contraire (règle 3.77).

$$réf(\sigma_1, \sigma_2) \supset (référence(\sigma_1) \wedge logique(\sigma_2)). \tag{3.75}$$

$$réf(\sigma_1, \sigma_2) \wedge réf\text{-}interne(\sigma_1) \supset \exists\sigma\, part\text{-}trans(\sigma, \sigma_1) \wedge part\text{-}trans(\sigma, \sigma_2). \tag{3.76}$$

$$réf(\sigma_1, \sigma_2) \wedge réf\text{-}externe(\sigma_1) \supset \neg(\exists\sigma\, part\text{-}trans(\sigma, \sigma_1) \wedge part\text{-}trans(\sigma, \sigma_2)). \tag{3.77}$$

Les *note*s sont aussi considérées comme des références, même si elles n'apparaissent pas dans le corps du discours proprement dit. Elles sont représentées en introduisant une ancre

de type *réf-note* qui représente le renvoi à la *note*. Imaginons par exemple une note de bas de page qui donne une brève définition de «*wapiti*»:

$réf\text{-}note(s_1)$.
$texte(s_1, \text{«}wapiti\text{»})$.
$note(s_2)$.
$texte(s_2, \text{«}grand\ cerf\ d'Amérique\ du\ Nord\text{»})$.
$réf(s_1, s_2)$.

Les types *réf-document* et *réf-note* ne peuvent référer qu'à des *doc*s et à des *note*s, respectivement.

$$réf(\sigma_1, \sigma_2) \wedge réf\text{-}document(\sigma_1) \supset doc(\sigma_2). \tag{3.78}$$
$$réf(\sigma_1, \sigma_2) \wedge réf\text{-}note(\sigma_1) \supset note(\sigma_2). \tag{3.79}$$

c) Certitude et alternatives logiques

Les relations *part* et *séq*, et *réf* peuvent être accompagnées d'une certitude. Toutefois, les mêmes réserves s'appliquent que pour la structure linguistique: il vaut mieux quand c'est possible utiliser les alternatives.

3.4.3 Structure de discours

Pour exprimer la structure de discours, nous reprenons les relations déjà vues à la section 2.3.2, à savoir *DSP* (*discourse segment purpose*), qui donne l'*intention* d'un segment de texte, et $\mathcal{D}om$ (dominance), qui donne la structure hiérarchique des *idées* d'un texte. Ces deux relations se notent *intention* et *domine* dans $\mathcal{L}$. Dans l'exemple ci-dessous, un élément σ domine un autre élément σ_s; ce dernier a pour intention le texte contenu dans un autre segment σ_i. On peut supposer que σ_s est une sous-section de σ, et que σ_i est un mot ou une expression qui résume bien l'idée exprimée dans σ_s.

$domine(\sigma, \sigma_s)$.
$intention(\sigma_s, \sigma_i)$.

La relation *domine* est très similaire à *part*: ses arguments sont de type *contenu-textuel* et elle respecte également les propriétés d'*antisymétrie*, d'*irréflexivité* et d'*injectivité*.

$$domine(\sigma_1, \sigma_2)[\mu] \supset (contenu\text{-}textuel(\sigma_1) \wedge contenu\text{-}textuel(\sigma_2)). \tag{3.80}$$
$$antisymétrique(domine). \tag{3.81}$$
$$irréflexive(domine). \tag{3.82}$$

$$injective(domine). \tag{3.83}$$

Nous définissons la relation *domine-trans*, qui est la contrepartie transitive de *domine*, comme suit:

$$\begin{aligned} domine(\sigma_1, \sigma_2) \quad &\supset \quad domine\text{-}trans(\sigma_1, \sigma_2). && (3.84)\\ & \quad\; transitive(domine\text{-}trans). && (3.85)\\ & \quad\; antisymétrique(domine\text{-}trans). && (3.86)\\ & \quad\; irréflexive(domine\text{-}trans). && (3.87) \end{aligned}$$

Quant au prédicat *intention*, il ne peut mettre en relation que des informations de type *contenu-textuel*, ce qui signifie que seules les intentions qui sont explicites dans le texte sont considérées ici.

$$intention(\sigma_1, \sigma_2)[\mu] \quad \supset \quad (contenu\text{-}textuel(\sigma_1) \wedge contenu\text{-}textuel(\sigma_2)). \tag{3.88}$$

La structure de discours est bien évidemment très difficile à déterminer en pratique, puisqu'a priori on ne dispose pas des idées de l'auteur. On peut cependant tirer avantage du fait que cette structure est généralement reflétée par la structure logique ainsi que par des indices linguistiques. Cette approche sera développée plus en détails au chapitre suivant.

3.5 Représentation des documents

Dans cette section nous voyons de façon concrète comment les divers types d'informations sont organisés dans la représentation des documents. Nous examinons d'abord de façon générale l'organisation des données hiérarchiques, qui forment la structure de base des documents, puis nous discutons de la représentation des informations non-hiérarchiques, et enfin nous concluons par un exemple complet de représentation de document.

3.5.1 Informations hiérarchiques

Les différentes parties d'un document se combinent suivant certaines «*règles*» de composition: par exemple, le titre d'une section apparaît toujours avant le corps du texte, la conclusion est toujours placée à la fin, etc. Ces règles sont très nombreuses, et sont de plus dépendantes du style de document: par exemple, pour un *article*, on combinera des *sections*, et les *remerciements* apparaîtront à la fin; alors que pour un *livre*, la structure de base est plutôt le *chapitre*, et les *remerciements* se trouvent au début.

Il ne s'agit pas ici de ré-écrire une grammaire universelle de documents, ce qui serait du reste inutile puisque ces règles de composition doivent être contrôlées à un niveau antérieur

au système de recherche d'informations, lors de l'écriture ou de la saisie du document. Nous nous contentons donc d'énoncer certaines règles générales s'appliquant à tout type de document, et qui sont à voir comme les *contraintes* sur les prédicats que nous avons définis dans les sections précédentes.

a) Divisions

Un document d est formé selon l'une des quatre possibilités suivantes: soit uniquement d'un *corps-doc*, soit d'un *début-doc* suivi d'un *corps-doc*, soit d'un *corps-doc* suivi d'un *fin-doc*, ou soit enfin de ces trois éléments combinés (règle 3.89).

$$
\begin{aligned}
&\mathit{document}(d) \supset \\
&\quad (\exists \sigma \ \mathit{part}(d,\sigma) \wedge \mathit{corps\text{-}doc}(\sigma)) \vee \\
&\quad (\exists \sigma_1, \sigma_2 \quad \mathit{part}(d,\sigma_1) \wedge \mathit{part}(d,\sigma_2) \wedge \\
&\qquad\qquad \mathit{début\text{-}doc}(\sigma_1) \wedge \mathit{corps\text{-}doc}(\sigma_2) \wedge \\
&\qquad\qquad \mathit{séq}(\sigma_1,\sigma_2)\ ) \vee \\
&\quad (\exists \sigma_1, \sigma_2 \quad \mathit{part}(d,\sigma_1) \wedge \mathit{part}(d,\sigma_2) \wedge \\
&\qquad\qquad \mathit{corps\text{-}doc}(\sigma_1) \wedge \mathit{fin\text{-}doc}(\sigma_2) \wedge \\
&\qquad\qquad \mathit{séq}(\sigma_1,\sigma_2)\ ) \vee \\
&\quad (\exists \sigma_1, \sigma_2, \sigma_3 \quad \mathit{part}(d,\sigma_1) \wedge \mathit{part}(d,\sigma_2) \wedge \mathit{part}(d,\sigma_3) \wedge \\
&\qquad\qquad \mathit{début\text{-}doc}(\sigma_1) \wedge \mathit{corps\text{-}doc}(\sigma_2) \wedge \mathit{fin\text{-}doc}(\sigma_3) \wedge \\
&\qquad\qquad \mathit{séq}(\sigma_1,\sigma_2) \wedge \mathit{séq}(\sigma_2,\sigma_3)\ )
\end{aligned}
\tag{3.89}
$$

Les *partie-doc* ont elles-mêmes pour contenu des *division*s ou des *document*s (règle 3.90).

$$
(\mathit{partie\text{-}doc}(\sigma_1) \wedge \mathit{part}(\sigma_1,\sigma_2)) \quad \supset \quad (\mathit{division}(\sigma_2) \vee \mathit{document}(\sigma_2)). \tag{3.90}
$$

b) Contenu des divisions

Le contenu d'une *division* est donné par une autre *division* ou par une composante de type *contenu-div* (règle 3.91).

$$
(\mathit{division}(\sigma_1) \wedge \mathit{part}(\sigma_1,\sigma_2)) \quad \supset \quad (\mathit{division}(\sigma_2) \vee \mathit{contenu\text{-}div}(\sigma_2)). \tag{3.91}
$$

Le *titre* ne peut apparaître que dans une *division*, une *figure* ou une *table* (règle 3.92). Il ne peut y avoir qu'un seul *titre* qui soit descendant direct d'un composant donné (règle

3.93). Enfin, le *titre* est forcément la première partie d'une *division*, ce qui n'est pas nécessairement le cas pour les *figures* ou les *tables* (règle 3.94).

$$(titre(\sigma_2) \wedge part(\sigma_1, \sigma_2)) \supset (division(\sigma_1) \vee figure(\sigma_1) \vee table(\sigma_1)). \quad (3.92)$$

$$(part(\sigma_1, \sigma_2) \wedge titre(\sigma_2) \wedge part(\sigma_1, \sigma_3) \wedge (\sigma_2 \neq \sigma_3)) \supset \neg titre(\sigma_3). \quad (3.93)$$

$$(division(\sigma) \wedge part(\sigma, \sigma_1) \wedge part(\sigma, \sigma_2) \wedge titre(\sigma_1) \wedge (\sigma_1 \neq \sigma_2)) \supset séq\text{-}trans(\sigma_1, \sigma_2). \quad (3.94)$$

c) Listes

Les *listes* sont des énumérations de *item-liste*. Supposons la liste l, qui énumère les pays membres de l'OTAN. On pourrait avoir deux des items de cette liste, i_1 et i_2, donnés comme suit:

$liste(l)$. $part(l, i_1)$. $part(l, i_2)$. $séq(i_1, i_2)$.

$item\text{-}liste(i_1)$. $item\text{-}liste(i_2)$. $texte(i_1, \text{«}\acute{E}tats\text{-}Unis\text{»})$. $texte(i_2, \text{«}Canada\text{»})$.

Les *listes* ne peuvent contenir que des *item-listes* (règle 3.95); et inversement, les *item-listes* ne peuvent être contenus que dans une *liste* (règle 3.96).

$$liste(\sigma_1) \wedge part(\sigma_1, \sigma_2) \supset item\text{-}liste(\sigma_2). \quad (3.95)$$

$$item\text{-}liste(\sigma_2) \wedge part(\sigma_1, \sigma_2) \supset liste(\sigma_1). \quad (3.96)$$

d) Figures

Les figures sont considérées comme des divisions logiques, qui, en plus de leur contenu, qui peut être de nature *contenu-textuel* ou *contenu-non-textuel* (il s'agit le plus souvent d'une image), peuvent comporter un *titre*, une *légende*, ou une *desc-fig*. Ces contraintes sont exprimées par la règle 3.97. Notons que les types *titre*, *légende*, et *desc-fig* n'apparaissent pas dans cette contrainte puisqu'ils sont déjà compris par *contenu-textuel*.

$$(figure(\sigma_1) \wedge part(\sigma_1, \sigma_2)) \supset (contenu\text{-}textuel(\sigma_2) \wedge contenu\text{-}non\text{-}textuel(\sigma_2)). \quad (3.97)$$

Les *desc-figs* ne peuvent apparaître que dans les *figures* (règle 3.98), quant aux *légendes*,

elles peuvent appararaître dans les *figures* ou les *tables* (règle 3.99).[10]

$$(\textit{desc-fig}(\sigma_2) \wedge \textit{part}(\sigma_1, \sigma_2)) \supset \textit{figure}(\sigma_1). \quad (3.98)$$

$$(\textit{légende}(\sigma_2) \wedge \textit{part}(\sigma_1, \sigma_2)) \supset (\textit{figure}(\sigma_1) \vee \textit{table}(\sigma_1)). \quad (3.99)$$

Soit par exemple la figure f, où apparaît un dessin représentant le *Père Noël*. Dans la représentation ci-dessous, l'image du Père Noël est représentée par f_i. Le titre de la figure, «*Figure 1*», est représenté par f_t. La légende, «*Le Père Noël en plein travail*», est donnée par f_l. Enfin, une description possible pour cette figure serait «*Un vieil homme joufflu à la barbe blanche, portant un grand sac sur les épaules*», donnée par f_d. La description est souvent utilisée pour l'affichage, lorsque l'image n'est pas disponible; cependant elle peut aussi servir pour l'indexation.

$\textit{figure}(f).$
$\textit{part}(f, i).$
$\textit{part}(f, f_t).$
$\textit{part}(f, f_l).$
$\textit{part}(f, f_d).$

$\textit{image-bitmap}(f_i).$
$\textit{titre}(f_t).$
$\textit{légende}(f_l).$
$\textit{desc-fig}(f_d).$
$\textit{texte}(f_t, \text{«}\textit{Figure 1}\text{»}).$
$\textit{texte}(f_l, \text{«}\textit{Le Père Noël...}\text{»}).$
$\textit{texte}(f_d, \text{«}\textit{Un vieil homme...}\text{»}).$

e) Tables

Les *tables*, tout comme les *figures*, sont considérées comme des divisions logiques, pouvant contenir un *titre*, une *légende*, ainsi que les *ligne-tables* formant le tableau. Ces contraintes sont définies à la règle 3.100.

$$(\textit{table}(\sigma_1) \wedge \textit{part}(\sigma_1, \sigma_2)) \supset (\textit{titre}(\sigma_2) \vee \textit{légende}(\sigma_2) \vee \textit{ligne-table}(\sigma_2)). \quad (3.100)$$

Les *ligne-tables* ne peuvent apparaître que dans une *table* (règle 3.101), à leur tour, elles ne peuvent contenir que des *cellule-tables* (règle 3.102). Inversement, les *cellule-tables* ne peuvent être contenues que dans une *ligne-table* (règle 3.103).

$$\textit{ligne-table}(\sigma_2) \wedge \textit{part}(\sigma_1, \sigma_2) \supset \textit{table}(\sigma_1). \quad (3.101)$$

$$\textit{ligne-table}(\sigma_1) \wedge \textit{part}(\sigma_1, \sigma_2) \supset \textit{cellule-table}(\sigma_2). \quad (3.102)$$

$$\textit{cellule-table}(\sigma_2) \wedge \textit{part}(\sigma_1, \sigma_2) \supset \textit{ligne-table}(\sigma_1). \quad (3.103)$$

Soit la table t ci-dessous, qui présente les thèmes principaux et secondaires pour un texte. Cette table est formée de deux lignes l_1 et l_2. Les cellules c_{11} et c_{12} sont les étiquettes

10. les contraintes sur le *titre* ont déjà été données par la règle 3.92

des colonnes, respectivement « *Thèmes principaux* » et « *Thèmes secondaires* ». Les cellules c_{21} et c_{22} représentent la première – et unique – ligne de valeurs, le thème principal étant « *rennes* », et les thèmes secondaires, « *Lapons, Père Noël* ».

$table(t)$.
$part(t, l_1)$.
$part(t, l_2)$.
$part(l_1, c_{11})$.
$part(l_1, c_{12})$.
$part(l_2, c_{21})$.
$part(l_1, c_{22})$.

$ligne\text{-}table(l_1)$.
$ligne\text{-}table(l_2)$.
$cellule\text{-}table(c_{11})$.
$cellule\text{-}table(c_{12})$.
$cellule\text{-}table(c_{21})$.
$cellule\text{-}table(c_{22})$.
$texte(c_{11}, \text{« } Thèmes\ principaux \text{ »})$.
$texte(c_{12}, \text{« } Thèmes\ secondaires \text{ »})$.
$texte(c_{21}, \text{« } rennes \text{ »})$.
$texte(c_{22}, \text{« } Lapons,Père\ Noël \text{ »})$.

3.5.2 Informations non-hiérarchiques

Si la structure de base d'un document est donnée par une hiérarchie, cette structure doit pouvoir être « *augmentée* » de façon à pouvoir accomoder certains types d'informations non-hiérarchiques [BBG+95]. Nous considérons dans $\mathcal{L}$ les informations non-hiérarchiques suivantes:

1° les informations supplémentaires ou non-séquentielles, comme par exemple une note en marge d'un texte ou l'auteur d'une citation;

2° les références à une autre partie du document ou à un autre document;

3° les informations ambiguës, comme par exemple un mot qui peut être *abbréviation* ou *terme technique*, ou un texte qui a plusieurs interprétations sémantiques;

4° les structures qui se chevauchent, comme par exemple un passage en langue étrangère qui commence dans un paragraphe et se termine dans un autre;

5° les structures parallèles, comme par exemple la traduction d'un passage de texte;

L'item n° 1 s'exprime dans $\mathcal{L}$ par les informations de *méta-contenu*, l'item n° 2, par la relation *réf*, les items n^{os} 3, 4 et 5 par les *alternatives*. La structure hiérarchique de base est donc conservée; ces informations peuvent être vues comme s'y ajoutant.

3.5.3 Un exemple de document

Nous proposons maintenant d'étudier un exemple complet de document dans $\mathcal{L}$. Pour ce faire, et afin de montrer également la correspondance entre notre langage de représentation

et d'autres standards pour la représentation et l'échange de documents, nous présentons d'abord le document à travers le formalisme TEI.

a) Représentation TEI

Le formalisme TEI (*Text Encoding Initiative*) [SM95b] [SM95a] consiste en un ensemble de recommandations pour l'encodage et l'échange de documents textuels. Ce formalisme est particulièrement adapté à notre problème de par sa richesse et sa flexibilité; de fait la majorité des informations de type *contenu-signifiant* dans $\mathcal{L}$ sont directement exprimables dans TEI, ce qui est loin d'être le cas avec d'autres formalismes comme HTML ou LaTeX.

L'ensemble des recommandations TEI est défini dans une grammaire ou DTD (pour *Document Type Definition*) SGML. TEI est donc basé sur le principe de marquage du texte: pour caractériser un élément du texte, on le précède et le fait suivre de *marques* (*tag*). Par exemple, pour identifier «*Représentation TEI*» comme le titre de cette section, on aurait:

<head id=ta>Représentation TEI</head>

Ici <head id=ta> représente la marque de début, et </head>, la marque de fin. Une information supplémentaire à propos de l'élément peut être ajoutée par le biais d'un *attribut*: il s'agit ici de id=ta, qui signifie que cet élément a pour identificateur ta.

La figure 3.15 donne un exemple de document TEI. Ce document consiste en une *entête électronique*, marquée par <teiheader>, et deux *divisions*, marquées par .

L'entête électronique fournit des informations externes au document, ce qui comprend mais ne se limite pas aux attributs bibliographiques.[11] Ici, deux informations sont spécifiées: le titre du document (<title>) et son auteur (<author>). Les marques <fileDesc> et <titleStmt> servent à identifier le contexte de ces informations, i.e. qu'il s'agit des attributs bibliographiques du document électronique.

Le *corps* du document est délimité par les marques <text> et <body>. Chaque division contient ensuite des paragraphes (<p>), qui contiennent à leur tour le texte. La première *division* est de type Introduction; elle consiste en deux *paragraphes*. La seconde *division* est un Chapitre ne contenant qu'un seul *paragraphe*, mais comportant aussi un *titre*, marqué par <head>.

L'incertitude peut s'exprimer dans TEI par la marque <certainty>. Dans notre exemple, on associe ainsi la certitude de .8 au *nom* «*Lapons*». L'attribut target renvoie à l'identificateur de l'objet visé, ici, p12. L'attribut degree spécifie la mesure de certitude, ici .8. Enfin, l'attribut locus permet de spécifier sur quelle partie doit porter l'incertitude; ici, par #gi (pour *general identifier*), on a précisé que l'incertitude porte sur la marque elle-même.

Les *alternatives* sont représentées dans TEI par l'élément <alt>, dont l'attribut targets

11. Il pourrait aussi s'agir d'informations sur le processus d'encodage par exemple.

<teiheader><fileDesc><titleStmt>
<title>Le renne de Laponie: Un animal méconnu</title>
<author>François Paradis</author>
</titleStmt></fileDesc></teiHeader>
<text><body>

Grand cervidé des terres du Nord, le renne demeure un animal mystérieux, de par son association au mythe du Père Noël, bien sûr, mais aussi à cause du secret qui entoure les <name id=p12>Lapons</name> <certainty target=p12 degree=.8 locus='#gi'>
, qui aujourd'hui encore en font l'élevage.
Nous discutons d'abord au <ref target=chap1>chapitre 1</ref> de la distribution géographique des rennes, pour ensuite examiner le rôle qu'il joue au sein de ces contrées reculées. .

<head>Où vivent les rennes?</head>
Les rennes <term id=p32a>Rangifer tarandus</term>
<term id=p32b>Tangifer tarandus</term><alt targets='p32a p32b'>
vivent dans les régions arctiques et sub-arctiques, principalement en Finlande, Suède, Norvège et Russie.

</body></text>

Figure 3.15. Exemple de document TEI

donne les pointeurs sur les alternatives et l'attribut weights leurs poids respectifs. Dans notre exemple, on a utilisé les *alternatives* pour donner deux orthographes possibles pour le nom scientifique de «*rennes*»: «*Rangifer tarandus*» (p32a) ou «*Tangifer tarandus*» (p32b).

L'exemple montre aussi l'usage de *noms* (<name>), *références* (<ref>) et *termes techniques* (<term>).

b) Représentation $\mathcal{L}$

La structure hiérarchique du document TEI donné à la section précédente demeure à peu de choses près la même lorsqu'exprimée dans $\mathcal{L}$, sauf pour ce qui est des *alternatives*. Cette structure est représentée schématiquement à la figure 3.16.

Le document d a pour corps de document la *corps-doc* cd, qui est composée des *divisions* *intro* et *chap1*. Ces divisions sont elles-mêmes composées des *paragraphes* p_1, p_2, et p_3. Les

éléments p_{11}, p_{12}, etc. représentent le contenu de ces paragraphes. Enfin, le *chapitre chap1* comporte aussi un titre, t_1.

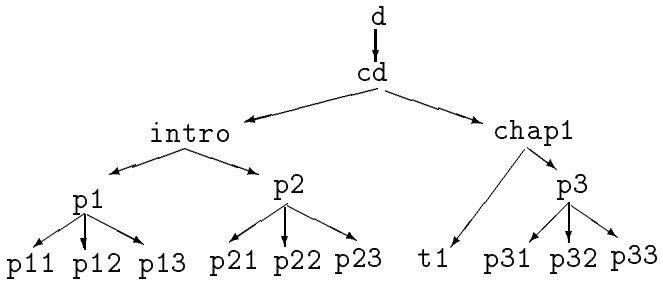


Figure 3.16. Structure du document d

Les éléments p_{32a} et p_{32b} ne sont pas placés en séquence à la suite de p_{31}, mais sont regroupés sous le *terme-technique* p_{32}, dont elles sont des *alternatives*.

La figure 3.17 donne la représentation complète de ce document dans $\mathcal{L}$. Notons qu'afin d'alléger cette représentation, les prédicats *contenu-textuel* ne sont pas exprimés. Ce type peut toujours être déduit du prédicat *texte* (règle 3.15).

L'entête électronique n'est pas représentée comme telle, mais au travers de certains attributs bibliographiques; ici, le *titre-doc* et l'*auteur*.

3.6 Conclusion

Nous avons dans ce chapitre défini le langage $\mathcal{L}$ pour la représentation des documents et de leurs index. Nous avons repris les types d'information définis au chapitre précédent, en montrant comment ils s'exprimaient dans les documents et comment les représenter dans $\mathcal{L}$, et avons également discuté d'autres aspects importants pour la recherche d'informations, notamment de la représentation de la certitude et des poids. Enfin, en définissant des règles sur les prédicats, nous avons levé toute ambiguïté quant aux contraintes qui sont supposées dans la représentation des documents.

L'emphase a été mise sur l'aspect descriptif du langage; le but n'étant pas de refaire une logique, mais bien de définir un formalisme approprié pour décrire les documents. Des critères comme la complétude ou l'efficacité sont donc complètement ignorés: il importait dans un premier temps de s'interroger sur ce qui doit être représenté par le langage avant de vérifier les caractéristiques du langage. D'ailleurs les problèmes qui pourraient survenir à ce niveau seraient forcément liés aux contraintes sur les prédicats, lesquelles ne seront pas forcément présentes dans un modèle opérationnel – elles sont surtout données ici pour lever toute ambiguïté quant au sens et à l'emploi des prédicats.

$document(d)$.
$titre\text{-}doc(d, \ll Le\ renne... \gg)$.
$auteur(d, \ll François\ Paradis \gg)$.
$part(d, cd)$.
$corps\text{-}doc(cd)$.
$part(cd, intro)$.
$part(cd, chap1)$.
$introduction(intro)$.
$chapitre(chap1)$.

$part(intro, p_1)$.
$part(intro, p_2)$.
$part(p_1, p_{11})$.
$part(p_1, p_{12})$.
$part(p_1, p_{13})$.
$part(p_2, p_{21})$.
$part(p_2, p_{22})$.
$part(p_2, p_{23})$.
$séq(intro, chap1)$.
$séq(p_1, p_2)$.
$séq(p_{11}, p_{12})$.
$séq(p_{12}, p_{13})$.
$séq(p_{21}, p_{22})$.
$séq(p_{22}, p_{23})$.
$réf(p_{22}, chap1)$.
$paragraphe(p_1)$.
$texte(p_{11}, \ll Grand...les \gg)$.
$nom(p_{12})[.8]$.
$texte(p_{12}, \ll Lapons \gg)$.
$texte(p_{13}, \ll ,qui...l'élevage. \gg)$.
$paragraphe(p_2)$.
$texte(p_{21}, \ll Nous...au \gg)$.
$réf\text{-}interne(p_{22})$.
$texte(p_{22}, \ll chapitre\ 1 \gg)$.
$texte(p_{23}, \ll de...reculées. \gg)$.

$part(chap1, p_3)$.
$part(chap1, p_{3t})$.
$part(p_3, p_{31})$.
$part(p_3, p_{32})$.
$part(p_3, p_{33})$.
$alternative(p_{32}, p_{32a})$.
$alternative(p_{32}, p_{32b})$.
$séq(p_{3t}, p_{31})$.
$séq(p_{31}, p_{32})$.
$séq(p_{32}, p_{33})$.
$titre(p_{3t})$.
$texte(p_{3t}, \ll Où\ vivent... \gg)$.
$paragraphe(p_3)$.
$texte(p_{31}, \ll Les\ rennes \gg)$.
$terme\text{-}technique(p_{32})$.
$texte(p_{32a}, \ll Rangifer... \gg)$.
$texte(p_{32b}, \ll Tangifer... \gg)$.
$texte(p_{33}, \ll vivent...Russie. \gg)$.

Figure 3.17. Exemple de représentation de document dans $\mathcal{L}$

Notre problématique peut aussi être vue comme la définition d'une ontologie, c'est-à-dire la formalisation des concepts et des relations entre les concepts pour un domaine donné, dans notre cas la représentation des documents textuels. D'ailleurs, le langage $\mathcal{L}$, est inspiré d'*Ontolingua* [Gru93, Gru92], un langage spécifiquement dédié à la définition des ontologies.

Le principal avantage de notre représentation est qu'elle lie dans un même formalisme le contenu textuel, attributs, sémantique, etc. Le fait de représenter l'index des documents au même niveau que les autres informations, nous permettra au chapitre suivant de définir des règles d'indexation pouvant déduire les thèmes à partir de la représentation sémantique, d'autres thèmes, du rôle ou de la fonction dans le texte, etc.

Enfin, puisque $\mathcal{L}$ est en fait un *méta-langage*, un langage de représentation opérationnel sera choisi lors de l'implémentation, selon les contraintes précises de l'application. Un

langage comme les graphes conceptuels permettrait de bien visualiser la structure des documents, un langage à base de frames pourrait être choisi pour son efficacité, etc. Ainsi, on peut consulter [Par94a] pour voir comment ces idées peuvent être exprimées avec les graphes conceptuels, ou [Par95a] avec *Ontolingua*. Il n'est pas exclu que plusieurs langages soient utilisés pour représenter différents types d'information, suivant leurs spécificités et les opérations qui doivent y être appliquées.

Chapitre 4

Comment dériver les thèmes

On dit que l'exception confirme la règle.
Elle la confirme en ce sens qu'elle lui donne
un soufflet.

Jules Tellier
(*Œuvres*)

Nous proposons dans ce chapitre de baser la dérivation des *thèmes* sur les théories de linguistique et de discours. Ainsi les mesures statistiques et anti-dictionnaires sont remplacés par des règles de dérivation qui nous semblent mieux approximer la notion intuitive qu'a l'utilisateur du *thème*. Les textes explicatifs sont particulièrement appropriés à notre approche, puisqu'ils sont généralement bien structurés et contiennent des expressions de *méta-discours* ou autres informations qui peuvent aider à dériver les thèmes.

Nous énonçons dans un premier temps les hypothèses qui serviront de base aux règles de dérivation, ce qui permet aussi de préciser la notion de thème dans notre modèle, et discutons de la mesure de représentatitivé du thème. Nous présentons les règles de dérivation, qui sont ensuite validées, d'abord par une étude de la fréquence d'apparition des mots-outils dans des collections tests, et ensuite par une expérience où des sujets devaient identifier les thèmes dans des expressions et dans un texte.

4.1 Généralités sur les thèmes

4.1.1 Hypothèses de dérivation

La dérivation de thèmes est un processus complexe, qui, nous semble-t-il, ne peut pas être approximé par une seule méthode de sélection comme le font les méthodes statistiques. Nous proposons diverses hypothèses quant à la nature du thème ou aux informations utilisées pour le dériver, hypothèses qui sont reprises par les règles de dérivation à la

section suivante. Ces hypothèses sont pour la plupart basées sur des théories linguistiques ou de discours. Cependant, puisqu'il est souvent difficile de les mettre en pratique, nous spécifions également comment les *approximer* en utilisant d'autres informations.

La première hypothèse concerne le contenu sémantique, qui stipule que le thème peut être déduit du sens *explicite*, c'est-à-dire qui apparaît dans le texte, ou *implicite*, c'est-à-dire déduit à partir de connaissances externes au document.

Hypothèse 1 (Contenu sémantique) *On peut déduire un thème à partir du contenu sémantique ou de connaissances sur le domaine.*

L'hypothèse de dépendance s'appuie sur les théories de grammaires de dépendances [vRW86]. Si une expression peut être décomposée en un élément *dominant* et un élément *dominé*, on peut considérer l'élément dominant comme thème pour cette expression. Bien que cette approche soit valable à tous les niveaux du texte, nous ne l'emploierons qu'au sein de la phrase.

Hypothèse 2 (Dépendance) *Dans une expression du type dominant-dominé, l'élément dominant est considéré comme un thème pour l'expression.*

Les trois prochaines hypothèses sont complémentaires: leur vision du thème est la même sauf qu'elles se situent à des niveaux différents: l'analyse statutaire est au niveau de la phrase, la progression thématique au niveau inter-phrase, et l'intention au niveau du discours. L'équivalence de ces notions est discutée dans [Car83].

Le thème de l'analyse statutaire correspond tout à fait à notre vision du thème en recherche d'informations. Il peut cependant s'avérer très difficile à déterminer en pratique. Une approximation du thème par les groupes nominaux, est souvent considérée suffisante pour les besoins de la recherche d'informations [BP86]. Cette approche est justifiée d'un point de vue linguistique par le fait que le thème – du moins du point de vue de Zemb [Mel87] – est toujours un groupe nominal: notons toutefois l'ajout considérable de bruit que cela entraîne puisque tous les groupes nominaux ne sont pas thèmes.

Hypothèse 3 (Analyse statutaire) *Chaque phrase peut être décomposée en thème, phème et rhème; c'est-à-dire ce dont on parle, ce qu'on en dit, et comment on le dit. Le thème peut être approximé par les groupes nominaux, ou par des éléments du vocabulaire qui sont accompagnés d'une marque sémantique.*

Notre hypothèse pour la progression thématique rejoint dans une certaine mesure les calculs de représentativités statistiques, puisqu'elle suppose qu'un élément qui est mentionné à plusieurs reprises dans une suite de phrases peut être considéré comme thème.

Hypothèse 4 (Progression thématique) *Le sujet ou le commentaire d'une phrase, lorsqu'il est répété dans une phrase ou un groupe de phrases subséquent, est considéré comme thème.*

Les intentions correspondent aux idées que l'auteur cherchait à exprimer dans son document: Grosz et Sidner suggèrent dans [GS86] qu'elle peuvent être assimilées aux thèmes. Encore une fois, puisque les intentions sont plutôt difficiles à déterminer en pratique, nous proposons plutôt de les approximer par le biais de certaines informations structurelles ou de méta-discours. Cette approche est justifiée dans [HL93].

Hypothèse 5 (Intentions) *Les intentions sont des thèmes; elles peuvent être approximées par certaines informations structurelles ou expressions-outils.*

Enfin, nous supposons que la structure logique du texte, c'est-à-dire sa décomposition en sections, reflète la structure de discours de l'auteur, laquelle peut être utilisée pour dériver les thèmes. Ceci paraît tout à fait raisonnable pour des textes explicatifs; si ce n'était pas le cas, les articles, livres, thèses, seraient incompréhensibles. Cette idée est aussi soutenue dans [HP93] et [CDB86].

Hypothèse 6 (Structure) *La structure logique du texte réflète la structure de discours et peut donc servir à dériver les thèmes.*

Dans un schéma de représentation de documents tel que $\mathcal{L}$, la décomposition du document en passages ne reflète pas toujours les passages au sens du lecteur, c'est-à-dire les sections, paragraphes, etc., mais peut être due aux limites ou particularités du langage de représentation. Ainsi, toutes les informations relatives au vocabulaire dans $\mathcal{L}$, comme par exemple *syntagme* ou *terme-technique*, sont traitées comme des passages au même titre qu'un paragraphe ou une section. Notre approche consiste à dériver les thèmes pour tous les passages, qu'ils soient motivés ou non par la structure logique, mais à ne conserver lors de l'implémentation que les passages situés au-delà d'un certain point dans la structure (voir l'expérimentation au chapitre suivant pour plus de détails). Ces passages sont dénommés les *passages d'indexation*. Cette approche est tout à fait similaire à l'emploi d'*unités d'indexation minimales*, tel que proposé dans [Ker84].

4.1.2 Mesure de représentativité

Nous reprenons ici la séparation entre représentativité *locale* et *globale*, telle que présentée à la section 1.2.2.

Nous proposons pour la représentativité locale d'utiliser un couplage avec les règles de dérivation, en associant $\text{reploc}(t,d)$ à la mesure du poids sur le *thème*, tel qu'il a été obtenu par une règle de dérivation:

$$\text{reploc}(t,d) = \begin{cases} \delta & \text{si } \textit{thème}(d,t)\{\delta\} \text{ existe} \\ 0 & \text{sinon} \end{cases}$$

Plutôt que de traduire l'importance relative du terme t au sein du passage d, reploc exprime la raison pour laquelle t a été sélectionné comme index de d, ce que nous dénommons par la suite la *justification*. On utilise pour exprimer la justification des symboles qui prennent des valeurs discrètes, dont l'ordre n'est pas défini a priori mais lors de l'application du jugement de pertinence.

Par exemple, $\text{reploc}(t, d) = H1$ signifie que t est un thème de d de par l'«hypothèse 1», et s'exprime dans $\mathcal{L}$ par $thème(d, t)\{H1\}$. Le symbole $H1$ est déduit lors de l'application de la règle correspondante.

La représentativité globale est associée au *contenu sémantique* du thème t, c'est-à-dire de l'apport du thème t par rapport à tous les thèmes du corpus. Nous reprenons pour ce faire la mesure de contenu sémantique inf, telle que définie au chapitre 2.

$$\text{repglob}(t, d) = inf(t)$$

Le contenu sémantique de t est alors donné par:

$$inf(t) = -\log m(t) = -\log\left(\frac{\text{fréquence} - \text{totale}(t)}{\text{taille} - \text{corpus}}\right) = \log\left(\frac{\text{taille} - \text{corpus}}{\text{fréquence} - \text{totale}(t)}\right)$$

où $m(t)$, la probabilité du thème t, est exprimée par le rapport entre le nombre d'occurrences de t dans la collection, $\text{fréquence} - \text{totale}(t)$, et la somme des occurrences de tous les thèmes dans le corpus, $\text{taille} - \text{corpus}$. Cette mesure tend vers 0 lorsque $\text{fréquence} - \text{totale}(t)$ est proche de $\text{taille} - \text{corpus}$, c'est-à-dire que le thème n'est pas assez *discriminant*, et tend vers $\log \text{taille} - \text{corpus}$ lorsque $\text{fréquence} - \text{totale}(t)$ est petit.

Enfin, la représentativité du terme t pour le passage d est donnée par le couple:

$$\text{rep}(t, d) = \langle \text{reploc}(t, d), \text{repglob}(t, d) \rangle$$

Rappelons qu'il s'agit d'un couple non homogène, où le premier élément, la justification, est un symbole, alors que le second, le contenu sémantique, est un nombre.

Cette mesure de représentativité permet un jugement de pertinence flexible qui peut varier selon divers facteurs, puisque l'ordre des justifications n'est pas fixé a priori. Ainsi, dans certains cas «*H5*» est préféré à «*H3*», alors que dans d'autres cas c'est l'inverse. Nous rediscuterons de ce point lors de la mise en œuvre du modèle au chapitre suivant.

4.2 Règles de dérivation

Les règles de dérivation de thèmes sont basées sur les hypothèses définies à la section 4.1.1. Elles sont exprimées à l'aide du langage $\mathcal{L}$[1], et sont donc similaires aux règles de

1. Pour une représentation alternative, plus intuitive mais aussi moins précise, on peut consulter [Par95c].

contraintes définies au chapitre précédent, sauf qu'elles ne font pas que vérifier des conditions syntaxiques ou sémantiques, mais produisent des faits. Cette nuance est indiquée par l'emploi du symbole «⇒» dans les règles de dérivation, plutôt que l'implication logique «⊃» qui était utilisée dans les règles de contraintes au chapitre précédent.

Un même thème peut être dérivé de différentes façons, en utilisant plusieurs règles; il existe alors plusieurs *justifications* au thème et seul le jugement de pertinence lors de la phase d'interrogation peut choisir la «*meilleure*» justification. C'est donc dire que la phase d'indexation doit dériver et conserver toutes ces alternatives.

Notons également que certaines règles décrites ici s'adressent à la linguistique, et peuvent donc être dépendantes de la langue. Nous le mentionnons quand c'est le cas, et donnons des exemples d'instanciation pour le français et l'anglais.

4.2.1 Contenu sémantique

Nous ne sommes pas tant intéressés dans notre modèle à la façon dont est représenté le contenu sémantique, mais surtout aux principes généraux qui permettent d'en déduire les thèmes. Les règles développées ici demeurent toutefois très générales: il n'est pas dans notre intention par exemple de modéliser le domaine.

Le contenu sémantique d'un élément de type *linguistique* – il peut s'agir d'expressions comme un groupe nominal, terme technique, etc. – est considéré comme thème pour cet élément. Ceci est exprimé par la règle 1.a: dans cette règle, on déduit que ς est thème de σ, sachant qu'il en exprime la sémantique et que σ est de type *linguistique*. Le poids $H1$ qui accompagne le prédicat *thème* représente la justification de ς, et fait référence aux hypothèses de dérivation définies à la section 4.1.1.

Règle 1.a (Contenu sémantique)

$$\text{sém}(\sigma, \varsigma) \wedge \text{linguistique}(\sigma) \quad \Rightarrow \quad \text{thème}(\sigma, \varsigma)\{H1\}.$$

Si deux éléments de contenu sémantique sont équivalents, et que l'un d'entre eux est thème de σ, alors le second est aussi thème de σ. Le thème ainsi dérivé a pour justification $H1$, et ce quel que soit le poids de la relation d'équivalence entre les deux thèmes (règle 1.b).

Règle 1.b (Équivalence)

$$\text{équivalent}(\varsigma_1, \varsigma_2)\{\delta_1\} \wedge \text{thème}(\sigma, \varsigma_1)\{\delta_2\} \quad \Rightarrow \quad \text{thème}(\sigma, \varsigma_2)\{H1\}.$$

En pratique, la synonymie conceptuelle est rare [Woo75], puisque si deux expressions sont jugées équivalentes, elles ont le même contenu sémantique. Plus courante est

l'*implication* entre concepts, c'est-à-dire quand on peut déduire un concept à partir d'un autre, mais pas réciproquement.

L'implication entre éléments de contenu sémantique est exprimée par le prédicat *implique*. Si un premier élément de contenu sémantique en implique un second, et que ce premier élément est thème de σ, alors le second l'est également (règle 1.c).

Règle 1.c (Inférence)

$$\text{implique}(\varsigma_1, \varsigma_2)\{\delta_1\} \wedge \text{thème}(\sigma, \varsigma_1)\{\delta_2\} \quad \Rightarrow \quad \text{thème}(\sigma, \varsigma_2)\{H1\}.$$

Par exemple, un document à propos du *Canada* peut aussi être considéré à propos de *pays*. Ainsi, si les faits suivants sont définis:

$implique(canada, pays).$
$thème(s, canada).$

alors on peut déduire $thème(s, pays)\{H1\}$.

4.2.2 Dépendance

Nous étudions dans cette section les liens de dépendance entre les constituants d'une phrase, et plus particulièrement au sein du syntagme nominal.

Si une expression σ consiste en deux éléments σ_1 et σ_2, et que σ_1 est dépendant de σ_2, alors le thème de σ_2 est aussi thème de σ (règle 2, première implication). La justification du thème dérivé ς_2 est $H2$.

Règle 2 (Dépendance)

$$\begin{aligned}
&(\text{dépendant}(\sigma_1, \sigma_2) \wedge \text{part}(\sigma, \sigma_1) \wedge \text{part}(\sigma, \sigma_2) \wedge \text{thème}(\sigma_2, \varsigma_2)\{\delta\}) \\
&\quad \Rightarrow \quad \text{thème}(\sigma, \varsigma_2)\{H2\}. \\
&(\text{dépendant}(\sigma_1, \sigma_2)\{\delta\} \wedge \text{part}(\sigma, \sigma_1) \wedge \text{part}(\sigma, \sigma_2) \wedge \\
&\qquad \text{thème}(\sigma_1, \varsigma_1)\{\delta_1\} \wedge \text{thème}(\sigma_2, \varsigma_2)\{\delta_2\}) \\
&\quad \Rightarrow \quad (\text{thème}(\sigma, \varsigma_2)\{H2\} \wedge \text{thème}(\sigma, \varsigma_1)\{\delta\}).
\end{aligned}$$

Lorsque la relation de dépendance entre σ_1 et σ_2 est seulement partielle, ce qui est indiqué par le fait qu'elle soit accompagnée d'un *poids*, alors bien entendu le thème de σ_2 demeure thème de σ, mais en plus, le thème de σ_1 est aussi thème de σ (règle 2, seconde implication). La justification du thème σ_1 est cependant moindre que celle de σ_2; elle est donnée par δ, qui correspond au poids sur la relation de dépendance.

La règle 2 indique comment dériver les thèmes à partir des liens de dépendances. Malheureusement ces liens ne sont généralement pas explicites dans le texte. Nous voyons maintenant comment déduire les liens de dépendance, soit à partir des catégories lexicales des constituants, soit à partir de mots-liens qui unissent les deux composants. Dans le premier cas, notre approche peut être assimilée à l'emploi d'un anti-dictionnaire, à ces deux différences près:

- il s'agit d'une sélection *contextuelle*, c'est-à-dire qu'un même mot peut être rejeté ou sélectionné dans des expressions différentes suivant son usage;

- il n'est pas impossible qu'un mot ou une expression rejetée de par la règle de dépendance soit sélectionné de par une autre règle.

En général, les constituants dépendants au sein d'un groupe nominal sont des *modificateurs*, c'est-à-dire des mots ou expressions qui n'ont pas d'existence propre en dehors du groupe nominal, et qui ne servent qu'à *modifier* ou à apporter des informations additionnelles au substantif principal. C'est le cas notamment pour les *déterminants* et les *adjectifs*.

Un substantif précédé d'un déterminant est toujours thème pour l'expression (règle 2.a). Par exemple, *livre* est thème pour «*le livre*», «*ce livre*», «*ton livre*», etc.

Règle 2.a (Déterminants)

$$(\text{déterminant}(\sigma_1) \wedge \text{substantif}(\sigma_2) \wedge \text{séq}(\sigma_1, \sigma_2)) \Rightarrow \text{dépendant}(\sigma_1, \sigma_2).$$

Dans $\mathcal{L}$ l'exemple «*le livre*» serait exprimé par les faits suivants:

$part(s, s1).$
$part(s, s2).$
$texte(s1, \text{«}le\text{»}).$
$texte(s2, \text{«}livre\text{»}).$
$thème(s2, livre).$
$séq(s1, s2).$
$déterminant(s1).$
$substantif(s2).$

Ces faits impliquent de par la règle 2.a le prédicat $dépendant(s1, s2)$. En appliquant la règle 2, on en déduit $thème(s, livre)\{H2\}$.

Nous ne faisons pas de traitement particulier pour les différents types d'*adjectif*s; qu'ils soient adjectifs qualificatifs ou épithètes, le nom auquel ils se rapportent est toujours thème

pour l'expression (règle 2.b). Par exemple, *livre* est thème pour «*livre rouge*».

Règle 2.b (Adjectifs)

$$(\text{adjectif}(\sigma_1) \wedge \text{substantif}(\sigma_2) \wedge \text{adjacent}(\sigma_1, \sigma_2)) \Rightarrow \text{dépendant}(\sigma_1, \sigma_2).$$

En anglais l'adjectif précède généralement le nom, et peut lui-même être précédé d'autres adjectifs, qui se placent alors selon un ordre très précis [Swa94, n°19]. Il peut cependant aussi se placer après le nom, comme dans «*ten years* old» (âgé *de dix ans*). En français, l'adjectif peut apparaître avant ou après le nom qu'il modifie, et peut même être accompagné d'un trait d'union, comme par exemple dans «*amour-propre*». Bien que ces différentes positions de l'adjectif traduisent une intention de transmettre un sens différent [Gre80, n°844], cette distinction est trop fine pour nos besoins, aussi nous contentons-nous à la règle 2.b de déclarer l'adjectif et le substantif comme *adjacents*.

Le substantif peut aussi être suivi d'un *adverbe*, qui agit alors comme un adjectif (règle 2.c). Par exemple, *paragraphe* est thème de l'expression «*le paragraphe* ci-dessus».

Règle 2.c (Adverbes utilisés comme adjectifs – Français)

$$(\text{adverbe}(\sigma_1) \wedge \text{substantif}(\sigma_2) \wedge \text{séq}(\sigma_1, \sigma_2)) \Rightarrow \text{dépendant}(\sigma_1, \sigma_2).$$

En anglais, le substantif peut être précédé d'un autre substantif, qui agit alors comme adjectif (règle 2.d). Le sens précis de cette forme est très variable [Swa94, n°424], mais le premier nom est toujours complément d'objet du second. Par exemple, l'expression «*iron bridge*» (*un pont en fer*) n'est pas à propos de *fer* mais bien de *pont*.

Règle 2.d (Substantifs utilisés comme adjectifs – Anglais)

$$(\text{substantif}(\sigma_1) \wedge \text{substantif}(\sigma_2) \wedge \text{séq}(\sigma_1, \sigma_2)) \Rightarrow \text{dépendant}(\sigma_1, \sigma_2).$$

Certains mots composés qui s'écrivent en un mot pourraient aussi être classés dans cette catégorie: «*toothbrush*» (*brosse à dents*) est à propos d'une *brosse*, «*housework*» (*travaux de ménage*) est à propos de *travaux*, etc. Ceci est aussi possible en français, comme dans «*portemanteau*», mais le plus souvent, les substantifs sont séparés par un trait d'union, comme dans «*oiseau-mouche*». Si le processus de marquage ne permet pas de séparer les substantifs dans les mots-composés, cette information peut être retrouvée au besoin grâce aux règles d'implication. Par exemple:

$$\textit{implique}(\textit{toothbrush}, \textit{brush}).$$

À l'inverse, dans certains mots composés, les composants pris isolément n'ont pas de sens. Par exemple, «*white elephant*» (littéralement, «*éléphant blanc*»; se dit d'une chose

inutile, superflue) est une figure de style; il n'est ni question de *blanc*, ni d'*éléphant*. De même, «*trou noir*», dans son interprétation astronomique, est à prendre comme un tout. Ces expressions constituent des exceptions aux règles 2.b à 2.e; elles doivent être prises en charge au niveau de la représentation du document, en ne permettant pas leur décomposition.

En français, l'apposition est «*un nom, ou un pronom, ou un infinitif, ou une proposition, qui se joint à un nom pour indiquer, comme le ferait un épithète, une qualité de l'être ou de l'objet dont il s'agit ... dans un sens plus large, elle ne sert parfois qu'à renforcer le nom*» [Gre80, n°341]. Deux substantifs qui se suivent sont forcément en apposition, et dans la plupart des cas, le second nom vient modifier le premier (règle 2.e). Par exemple, dans «*roi soleil*» il est question d'un *roi* (Louis XIV), mais pas du *soleil*.

Règle 2.e (Substantifs en apposition – Français)

$$(\text{substantif}(\sigma_1) \wedge \text{substantif}(\sigma_2) \wedge \text{séq}(\sigma_1, \sigma_2)) \Rightarrow \text{dépendant}(\sigma_2, \sigma_1).$$

Il existe plusieurs exceptions à cette règle, mais ce sont généralement des figures de style qu'on a peu de chance de retrouver dans un texte explicatif. Ainsi, «*maître corbeau*» est à propos d'un *corbeau*, «*fin décembre*», ellipse de «*fin de décembre*» est à propos de *décembre*.

Plusieurs expressions comportant un nom propre sont considérées comme des appositions: «*fusée Ariane*» et «*fleuve St-Laurent*» sont à propos d'une *fusée* et d'un *fleuve*, respectivement. Le nom propre – ici «*Ariane*» et «*St-Laurent*» – étant de type *nom*, il est aussi dérivé comme thème, mais de par la règle 3.c. Ceci est nécessaire pour résoudre les ellipses: «*Ariane*» au lieu de la «*fusée Ariane*», «*le St-Laurent*» au lieu du «*fleuve St-Laurent*».

La similitude entre les règles 2.d et 2.e est trompeuse: dans les deux cas on traite effectivement de substantifs qui se suivent, mais pour l'anglais, le second est porteur de thème, tandis que pour le français, c'est le premier.

Le *génitif* est employé en anglais pour marquer la possession, les traits physiques, les mesures, etc. [Swa94, n°261]. Il est formé en ajoutant «*'s*» au singulier ou «*'*» au pluriel des substantifs. Dans la forme *génitive*, bien que le substantif modifié demeure le thème principal, l'expression qui vient le modifier a une certaine «*existence*». Ceci est traduit dans la règle 2.f par une relation de dépendance *partielle* entre σ_1 et σ_3; bien que σ_1 soit toujours dépendant de σ_3, il l'est dans une moindre mesure puisqu'il a aussi son existence propre.

Règle 2.f (Génitif – Anglais)

$$(\text{séq}(\sigma_1, \sigma_2) \wedge \text{séq}(\sigma_2, \sigma_3) \wedge (\text{texte}(\sigma_2, \text{«'s»}) \vee \text{texte}(\sigma_2, \text{«'»}))$$
$$\Rightarrow \quad \text{dépendant}(\sigma_1, \sigma_3)\{H2\text{-génitif}\}.$$

Comparons par exemple «*dog food*» (*nourriture pour chiens*) avec «*the dog's food*» (*la nourriture du chien*). Dans le premier cas, *chiens* est un concept générique, qui vient classifier le type de nourriture. Dans le second cas, *chien* fait référence à une bête en particulier, qui a une «*existence*» dans le discours; par conséquent, *chien* est thème pour «*the dog's food*» avec la justification *H2-génitif* (*nourriture* demeure bien entendu le thème principal de cette expression).

La forme du génitif qui n'est pas suivie de nom [Swa94, n°263], comme dans «*Mary is at the hairdresser's*» (*Marie est chez le coiffeur*) n'est pas traitée par nos règles.

Les groupes prépositionnels sont aussi porteurs de beaucoup d'information dans les syntagmes nominaux. Malheureusement, leur sens s'avère très difficile à déterminer en pratique, puisqu'il peut changer pour une même préposition suivant le contexte. Selon le cas, c'est le premier ou le second élément qui domine: ainsi, une «*réunion d'information*» est à propos d'une *réunion*, tandis qu'un «*drôle d'animal*» est à propos d'un *animal*. Quand bien même l'élément dépendant peut être identifié, dans certains cas l'élément dominant ne correspond pas au thème. Par exemple, l'expression «*le fond de la tasse*» est plus à propos de *tasse* que de *fond*, malgré la relation de dépendance inverse. De même, dans «*recherche d'informations*», «*recherche*» et «*informations*» sont difficilement dissociables. Enfin, dans bon nombre de cas, les deux constituants peuvent être thèmes de l'expression. Par exemple, dans «*pollution du Saint-Laurent*», *pollution* et, dans une moindre mesure, *Saint-Laurent*, sont thèmes.

À défaut de pouvoir utiliser les groupes prépositionnels, nous dérivons le thème dans les propositions relatives. Les propositions relatives sont introduites par un pronom relatif («*qui*», «*que*», «*quoi*», «*dont*», etc. en français; «*who*», «*whom*», «*which*», «*that*», etc. en anglais) ou un adverbe relatif («*où*»). En français, bien qu'elles puissent exprimer le but, l'opposition, l'hypothèse, etc., elles ont toujours valeur propre d'adjectif [Gre80, n°2607]. C'est pourquoi nous supposons que le substantif auquel se rattache la proposition relative est thème (règle 2.g).

Règle 2.g (Propositions relatives)

$$\begin{aligned}&(\text{substantif}(\sigma_1) \wedge (\text{pronom-relatif}(\sigma_2) \vee \text{adverbe-relatif}(\sigma_2)) \wedge \\ &\text{séq}(\sigma_1, \sigma_2) \wedge \text{séq}(\sigma_2, \sigma_3)) \\ &\quad \Rightarrow \quad \text{dépendant}(\sigma_3, \sigma_1)\{H2\text{-relative}\}.\end{aligned}$$

En anglais, on distingue les propositions relatives qui *identifient* le nom, c'est-à-dire qui sont indispensables pour sa compréhension, de celles qui ne l'identifient pas, c'est-à-dire qui ne sont pas nécessaires pour sa compréhension. Par exemple, dans l'expression «*Daniel, who borrowed my book*» (*Daniel, qui a emprunté mon livre*), même en éliminant la clause relative, on sait toujours de quel *Daniel* il est question. Par contre, dans «*the man who borrowed my book*» (*l'homme qui a emprunté mon livre*), la clause relative est nécessaire pour identifier l'homme en question.

Ces deux types de propositions relatives induisent des poids différents sur la relation de dépendance. Le thème d'une proposition relative *identificatrice* est déduit avec une justification égale à *H2-relative*, tandis que pour une proposition *non-identificatrice*, la justification est *H2-relative-non-id*. Puisque dans ce second cas, le thème est plutôt une information additionnelle de moindre importance par rapport au nom auquel la proposition se rapporte, on a toujours: $H2\text{-}relative\text{-}non\text{-}id \leq H2\text{-}relative$.

Les propositions *non-identificatrices* sont toujours séparées du reste de la phrase par une pause, ce qui se traduit dans le texte par une virgule [Swa94, n°]. Ainsi, la règle 2.h est similaire à la règle 2.g, si ce n'est que le pronom ou l'adverbe relatif est précédé d'une virgule. Cette règle est aussi applicable au français.

Règle 2.h (Propositions relatives non-identificatrices)

$$\begin{aligned}&(\text{substantif}(\sigma_1) \wedge \text{texte}(\sigma_2, «,») \wedge (\text{pronom-relatif}(\sigma_3) \vee \text{adverbe-relatif}(\sigma_3)) \wedge \\ &\text{séq}(\sigma_1, \sigma_2) \wedge \text{séq}(\sigma_2, \sigma_3) \wedge \text{séq}(\sigma_3, \sigma_4)) \\ &\quad \Rightarrow \quad \text{dépendant}(\sigma_4, \sigma_1)\{H2\text{-relative-non-id}\}.\end{aligned}$$

Afin de dériver les relations de dépendance entre des éléments plus complexes, comme par exemple une expression « *déterminant – adjectif – substantif* » il suffit de savoir dériver les groupes nominaux, que nous considérons comme des substantifs, et de continuer à dériver les relations de dépendances deux-à-deux.

Un segment de contenu textuel σ constitué de deux sous-parties dont l'une est dépendante de l'autre, est considéré comme un *substantif* [2] (règle 2.i).

Règle 2.i (Substantif)

$$(\text{part}(\sigma, \sigma_1) \wedge \text{part}(\sigma, \sigma_2) \wedge \text{dépendant}(\sigma_1, \sigma_2)) \Rightarrow \text{substantif}(\sigma).$$

De cette façon, voici la séquence de règles applicables à l'expression «*le livre rouge que j'ai acheté*»:

1° la règle de l'adjectif (2.b), identifiant la relation de dépendance entre «*livre*» et «*rouge*»;

2° la règle du substantif (2.i), formant le groupe nominal «*livre rouge*»;

3° la règle du déterminant (2.a) entre «*le*» et «*livre rouge*»;

4° la règle du substantif (2.i), pour former «*le livre rouge*»;

5° enfin, la règle de proposition relative (2.g) entre «*le livre rouge*» et «*j'ai acheté*».

Les thèmes suivants seraient ainsi déduits: *livre*, avec la justification *H2*, *livre rouge*, avec la justification *H2*, et *le livre rouge*, avec la justification *H2-relative*.

2. Il s'agit plus précisément d'un groupe nominal.

4.2.3 Analyse statutaire

Les éléments identifiés comme *marque-thème*, c'est-à-dire qui sont considérés comme des thèmes au sens de l'analyse statutaire, sont des thèmes pour la phrase dans laquelle ils apparaissent (règle 3).

Règle 3 (Thème)

$$\text{marque-thème}(\sigma) \wedge \text{thème}(\sigma, \varsigma)\{\delta\} \wedge \text{part-trans}(\sigma_p, \sigma) \wedge \text{phrase}(\sigma_p) \Rightarrow \text{thème}(\sigma_p, \varsigma)\{H3\}.$$

Nous avons vu à la section 2.3.1 comment la séparation d'une phrase en *thème*, *phème*, et *rhème* pouvait être difficile à déterminer en pratique. Nous proposons dans la suite de cette section des règles pour approximer cette information.

On peut approximer le thème par les types *index* et *syntagme* (règles 3.a et 3.b). Les *index* sont de très bonnes indications du thème, puisqu'ils ont été manuellement sélectionnés comme des sujets de recherche possible (recherche qu'effectue alors le lecteur à l'aide de l'index imprimé à la fin du document). Les *syntagme*s sont moins précis: bien que tout thème soit *syntagme*, il existe des *syntagme*s qui appartiennent au rhème.

Règle 3.a (Index)

$$\text{index}(\sigma) \Rightarrow \text{marque-thème}(\sigma).$$

Règle 3.b (Syntagme nominal)

$$\text{syntagme}(\sigma) \Rightarrow \text{marque-thème}(\sigma).$$

Une autre possibilité consiste à utiliser les marques sémantiques du contenu textuel. Ceci est justifié par le fait que ces marques sont généralement ajoutées manuellement par l'encodeur, dans le but de discriminer le texte pour des raisons *sémantiques*, par opposition aux marques qui ne servent qu'à en spécifier l'apparence *physique* (ceci n'empêchant pas bien entendu le texte identifié par des marques sémantiques d'avoir également une apparence physique qui le distingue du texte qui l'entoure) .

La règle 3.c montre les divers types d'information qui peuvent approximer le thème. Chacune de ces implications dérive un thème avec une justification $H3$-α, où α est un identificateur du type sémantique. Par exemple, une *glose* est thème pour la phrase où elle apparaît avec une justification de $H3$-*glose*.

Règle 3.c (Marques sémantiques)

$$\alpha(\sigma) \wedge \text{part-trans}(\sigma_p, \sigma) \wedge \text{phrase}(\sigma_p) \wedge \text{thème}(\sigma, \varsigma)\{\delta\} \wedge$$

$(\alpha = \text{terme-technique} \vee \alpha = \text{nom} \vee \alpha = \text{code} \vee \alpha = \text{exemple} \vee \alpha = \text{glose} \vee$
$\alpha = \text{accentué} \vee \alpha = \text{réf-document} \vee \alpha = \text{item-étiquette} \vee \alpha = \text{ident-marque} \vee$
$\alpha = \text{ident-marque-déf} \vee \alpha = \text{ident-attr} \vee \alpha = \text{ident} \vee \alpha = \text{abbréviation} \vee$
$\alpha = \text{expansion})$
$\Rightarrow \quad \text{thème}(\sigma_p, \varsigma)\{H3\text{-}\alpha\}.$

4.2.4 Progression thématique

Les éléments identifiés comme *marque-sujet*, c'est-à-dire les *sujets* par opposition aux *commentaires*, sont des thèmes pour la phrase dans laquelle ils apparaissent (règle 4.a).

Règle 4.a (Sujet)

$\text{marque-sujet}(\sigma) \wedge \text{part-trans}(\sigma_p, \sigma) \wedge \text{phrase}(\sigma_p) \wedge \text{thème}(\sigma, \varsigma)\{\delta\}$
$\Rightarrow \quad \text{thème}(\sigma_p, \varsigma)\{H4\}.$

Les règles 4.b et 4.c traitent des relations inter-phrases. Elles peuvent se résumer comme suit: le sujet ou le commentaire d'une phrase répété dans une phrase subséquente est thème pour les deux phrases. Nous généralisons ce principe en considérant des segments de texte quelconques, et en remplaçant le sujet par le thème.

La règle 4.b se lit comme suit: si le texte τ associé au thème ς d'un passage σ_{p1} se retrouve dans un passage σ_{p2} suivant immédiatement σ_{p1}, alors ς est thème de $\sigma_{p1\cdot 2}$, le passage formé de σ_{p1} et σ_{p2}.

Règle 4.b (Répétition du thème)

$(\text{thème}(\sigma_{p1}, \varsigma)\{\delta_1\} \wedge \text{thème}(\sigma, \varsigma)\{\delta_2\} \wedge \text{part-trans}(\sigma_{p1}, \sigma) \wedge \text{texte}(\sigma, \tau) \wedge$
$\text{séq}(\sigma_{p1}, \sigma_{p2}) \wedge \text{contient-texte}(\sigma_{p2}, \tau) \wedge$
$\text{part}(\sigma_{p1\cdot 2}, \sigma_{p1}) \wedge \text{part}(\sigma_{p1\cdot 2}, \sigma_{p2}))$
$\Rightarrow \quad \text{thème}(\sigma_{p1\cdot 2}, \varsigma)\{H4\}.$

La règle 4.c est similaire, si ce n'est que cette fois on s'intéresse à la répétition du texte d'une *marque-commentaire*.

Règle 4.c (Répétition du commentaire)

$(\text{marque-commentaire}(\sigma) \wedge \text{thème}(\sigma, \varsigma)\{\delta\} \wedge \text{part-trans}(\sigma_{p1}, \sigma)\text{texte}(\sigma, \tau) \wedge$
$\text{séq}(\sigma_{p1}, \sigma_{p2}) \wedge \text{contient-texte}(\sigma_{p2}, \tau) \wedge$
$\text{part}(\sigma_{p1\cdot 2}, \sigma_{p1}) \wedge \text{part}(\sigma_{p1\cdot 2}, \sigma_{p2}))$
$\Rightarrow \quad \text{thème}(\sigma_{p1\cdot 2}, \varsigma)\{H4\}.$

4.2.5 Intentions

Les thèmes d'un passage identifié comme l'*intention* d'un autre passage σ_s, sont aussi des thèmes pour σ_s (règle 5).

Règle 5 (Intentions)

$$(\text{thème}(\sigma, \varsigma)\{\delta\} \wedge \text{intention}(\sigma_s, \sigma)) \;\Rightarrow\; \text{thème}(\sigma_s, \varsigma)\{H5\}.$$

Cette information n'est pour ainsi dire jamais présente explicitement dans les documents; par contre, et ceci est particulièrement vrai dans les textes explicatifs, certains indices présents dans le texte peuvent permettre de la déduire.

Ainsi, les titres de sections sont généralement de bons indicateurs de l'intention de l'auteur, puisqu'ils résument ce dont il est question dans la section correspondante. Par exemple, le titre de ce chapitre, «*Comment dériver les thèmes*», annonce d'emblée qu'il est question de la dérivation des thèmes.

Le titre σ d'une section σ_s, pour autant qu'il ne s'agit pas d'une *localisation*, est donc considéré comme décrivant son *intention* (règle 5.a). On identifie comme *localisation* les *titres-outils*, c'est-à-dire les titres qui ne font que décrire le *type* de la section sans être porteurs d'information en eux-mêmes: ce sont les titres dont le texte consiste en «*Introduction*», «*Conclusion*», etc.[3]

Règle 5.a (Titres)

$$(\text{titre}(\sigma) \wedge \text{part}(\sigma_s, \sigma) \wedge \neg\text{localisation}(\sigma)) \;\Rightarrow\; \text{intention}(\sigma_s, \sigma).$$

Dans la littérature scientifique, l'auteur annonce souvent la structure de son discours par le biais de mots ou d'expressions particulières. Nous qualifions ces expressions de *méta-discours* , ainsi nommées parce qu'il s'agit de ce que l'auteur dit au sujet de son discours. Voici quelques exemples de *méta-discours*: «*nous discutons maintenant de...* », «*la figure x montre...* », «*le chapitre suivant décrit...* », etc.

Ces expressions de *méta-discours* sont similaires aux expressions dites «*cue phrases*», qui sont étudiées dans [HL93]. Cependant, alors que les *cue phrases* sont plutôt des mots-liens qui indiquent les limites entre les segments de discours, les expressions de méta-discours, au sens où nous l'entendons, indiquent directement l'intention pour un segment de discours.

Nous nous intéressons aux expressions du méta-discours qui *décrivent* un segment de texte. Une expression de méta-discours est décomposée en trois éléments, pouvant apparaître dans un ordre quelconque: un *verbe-descriptif*, qui précise le type de description, une *localisation*, qui réfère au segment de discours en question, et enfin, l'intention elle-même.

3. Les titres-outils ne sont identifiés que sur la base du texte qui leur est associé: il est tout à fait possible qu'une section de type *introduction* ait un titre significatif, si le texte diffère de «*Introduction*». Une liste plus conséquente de titres-outils est donnée plus bas.

La grammaire présentée à la figure 4.1 définit les *verbe-descriptifs* et *localisations* considérés pour le français. Il s'agit bien entendu d'une liste non limitative. Les verbes ne sont donnés qu'à l'infinitif dans la grammaire; ils apparaîtront conjugués dans les expressions (par exemple, «*démontre*» au lieu de «*démontrer*»), accompagnés ou non d'un pronom (par exemple, «*nous démontrons*»), et à la forme active ou passive (par exemple, «*il est démontré*»). Les accords en genre ou en nombre sont délibérément omis de cette grammaire afin d'en simplifier l'expression.

`localisation` ::= [`dét`] `dés` [`ordre`] | `ordre-ind` | 'ici' | 'maintenant'
`dét` ::= `dét1` | ['dans'] `dét2`
`dét1` ::= 'au' | 'à la'
`dét2` ::= 'ce' | 'cet' | 'cette' | 'le' | 'la' | 'l' | 'notre'
`ordre` ::= 'suivant' ['e'] | 'précédent' ['e'] | 'qui suit' | 'qui précède'
| `ordre-ind` | nombre
`ordre-ind` ::= 'ci-dessus' | 'ci-dessous' | 'ci-après'
`dés` ::= 'figure' | 'graphe' | 'graphique' | 'illustration' | 'tableau' | 'table' |
'paragraphe' | 'exemple' | 'article' | 'papier' | 'ouvrage' | 'section'
| 'chapitre' | 'livre' | 'thèse' | 'rapport' | 'travail' | 'approche' |
'problématique' | 'introduction' | 'conclusion' | 'bibliographie' |
'remerciements' | 'références' | 'index' | 'résumé' | 'annexe'
`verbe-descriptif` ::= présenter | décrire | montrer | expliquer | démontrer |
illustrer | prouver | concerner | développer | énoncer |
résumer | aborder | consister | exposer | examiner

Figure 4.1. Quelques expressions de méta-discours en français

---

Voici quelques exemples d'expressions de méta-discours produites en utilisant ces définitions: «*nous décrivons à la section suivante...*», «*la figure 1.2 montre...*», «*nous abordons maintenant...*», «*l'exemple ci-dessous illustre...*», «*notre travail consiste...*», etc.

Soit par exemple l'expression «*Nous discutons à la section 1.2.2 de représentativité*», dont la représentation dans $\mathcal{L}$ est donnée ci-dessous.

$texte(p_v, \text{«}Nous\ discutons\text{»})$.
$part(p_l, p_{l1})$.
$part(p_l, p_{l2})$.
$texte(p_{l1}, \text{«}à\ la\text{»})$.
$texte(p_{l2}, \text{«}section\ 1.2.2\text{»})$.
$texte(p_i, \text{«}de\ représentativité\text{»})$.
$séq(p_v, p_l)$.
$séq(p_l, p_i)$.
$séq(p_{l1}, p_{l2})$.
$verbe\text{-}descriptif(p_v)$.
$localisation(p_l)$.
$référence(p_{l2})$.
$réf(p_{l2}, s_{1.2.2})$.

Dans cet exemple, le *verbe-descriptif* est donné par p_v, la *localisation* par p_l, et l'intention, par p_i. Notons que p_l est lui-même composé de deux éléments, dont l'un est une *référence* à la section 1.2.2. Grâce à cette information, on pourra déduire que p_i est une intention de $s_{1.2.2}$ (règles 5.b et 5.c).

La règle 5.b permet de dériver les intentions d'une section σ à partir d'une expression de méta-discours formée d'un *verbe-descriptif* σ_v, d'une *localisation* σ_l, et d'une *intention*, σ_i.

Règle 5.b (Méta-discours)

$$\begin{aligned}&(\text{verbe-descriptif}(\sigma_v) \wedge \text{localisation}(\sigma_l) \wedge \text{intention}(\sigma_i) \wedge \\ &\text{adjacent}(\sigma_v, \sigma_l) \wedge \text{adjacent}(\sigma_l, \sigma_i) \wedge \text{désigne}(\sigma_l, \sigma)) \\ &\quad \Rightarrow \quad \text{intention}(\sigma, \sigma_i).\end{aligned}$$

Les trois éléments σ_v, σ_l et σ_i sont liés par la relation *adjacent*, qui rappelons-le, est transitive, ce qui signifie que les trois éléments peuvent apparaître dans un ordre quelconque, pourvu qu'ils se succèdent. Dans notre exemple, au lieu de «*Nous discutons à la section 1.2.2 de représentativité*», on pourrait avoir: «*Nous discutons de représentativité à la section 1.2.2*», «*À la section 1.2.2, nous discutons de représentativité*», «*À la section 1.2.2, la représentativité est discutée*», ou enfin, «*La représentativité est discutée à la section 1.2.2*».[4]

Le prédicat *désigne* sert à lier une expression de localisation à l'entité qu'elle dénote. Cette information peut être déduite de la *localisation* si cette dernière contient une référence à l'entité (règle 5.c): ainsi pour l'exemple ci-dessus, *désigne*$(p_l, s_{1.2.2})$est déduite.

Règle 5.c (Désignation)

$$\begin{aligned}&(\text{localisation}(\sigma) \wedge \text{part-trans}(\sigma, \sigma_p) \wedge \text{référence}(\sigma_p) \wedge \text{réf}(\sigma_p, \sigma_r)) \\ &\quad \Rightarrow \quad \text{désigne}(\sigma, \sigma_r).\end{aligned}$$

Lorsque les références ne sont pas disponibles pour une *localisation*, il faut faire appel à une analyse plus poussée: des expressions comme «*ici*» ou «*maintenant*» font généralement référence à la section courante, une expression de genre «*la section ci-dessus*» fait appel à la section précédente, etc.

Il va sans dire que ces règles pour le traitement des expressions de méta-discours sont très incomplètes; nous considérons cependant qu'elles recouvrent la vaste majorité de telles expressions dans les textes explicatifs. Voici d'autres formes qui pourraient être traitées:

- l'omission de la *localisation*, comme par exemple «*nous démontrons la complétude du langage*», qui réfère habituellement au passage courant;

4. La dernière combinaison, σ_d-σ_l-σ_v, ne nous semble pas très usitée, mais elle est demeure permise par la règle 5.b.

- les *localisations* et/ou *intentions* multiples, comme par exemple, «*nous discutons aux sections 3.4.1 et 3.4.2 de la structure linguistique et logique, respectivement*»;
- les intentions coupées en deux dans l'expression, comme par exemple «*l'usage de l'incertitude est illustré pour quelques prédicats dans l'exemple ci-dessous*»;
- les intentions au premier degré, du genre «*le but de notre travail consiste...*» ou «*l'intérêt de notre approche réside...*».

4.2.6 Structure de discours

L'organisation hiérarchique du document joue un rôle important dans l'identification des thèmes. Non seulement la prise en compte de cette structure est-elle souvent nécessaire à la bonne compréhension du document, mais de plus, elle reflète l'importance relative des thèmes. La division en *chapitre/section/sous-section* de cette thèse, par exemple, reflète son organisation en thèmes et sous-thèmes, chaque sous-section précisant davantage le thème global ou apportant de nouveaux thèmes (mais qui sont toujours secondaires par rapport au thème principal).

Notre première supposition est que la structure de discours de l'auteur, c'est-à-dire l'organisation hiérarchique de ses idées, est réflétée dans le texte par la *structure logique*, c'est-à-dire la décomposition en *division*. Ainsi, un lien *part* entre deux *divisions* permet de dériver le lien *domine* (règle 6.a).

Règle 6.a (Structure de discours)

$$(\text{part}(\sigma_1, \sigma_2) \wedge \text{division}(\sigma_1) \wedge \text{division}(\sigma_2)) \Rightarrow \text{domine}(\sigma_1, \sigma_2).$$

La structure de discours ne se limite pas à ces relations: elle peut aussi inclure des liens entre paragraphes ou à l'intérieur d'un même paragraphe. D'autres travaux se sont intéressés à ce type de liens [MH91] [Hea94]. Nous pensons cependant qu'il n'est pas utile d'avoir un tel degré de finesse si les segments de discours considérés sont plus petits que les passages «*minimaux*» visualisés par l'utilisateur.

La règle 6.b reprend l'idée de la stratégie de remontée de termes, en ajoutant la contrainte suivante: un thème n'est propagé à sa section-mère que s'il apparaît dans chacune de ses sections-filles.

Règle 6.b (Héritage ascendant)

$$(\forall \sigma_i \ \text{domine}(\sigma, \sigma_i) \supset \text{thème}(\sigma_i, \varsigma)\{\delta\}) \quad \Rightarrow \quad \text{thème}(\sigma, \varsigma)\{H6\}.$$

L'*introduction* et le *résumé* sont des cas particuliers de cette règle, où tous les thèmes de la division sont propagés à la division-mère (règle 6.c), qu'ils apparaissent ou non dans

les autres divisions-filles. En effet, les divisions de ce type ont pour but de présenter la problématique, et donc de préciser le thème principal de la division-mère, et/ou de présenter l'organisation de la division-mère, et donc les thèmes des autres divisions-filles.

Règle 6.c (Introduction, Résumé)

$$((\text{introduction}(\sigma_i) \vee \text{résumé}(\sigma_i)) \\ \wedge \text{thème}(\sigma_i, \varsigma)\{\delta\} \wedge \text{part}(\sigma, \sigma_i)) \quad \Rightarrow \quad \text{thème}(\sigma, \varsigma)\{H6\}.$$

Ce principe est aussi applicable à la *conclusion*, mais dans une moindre mesure: les thèmes sont ici dérivés avec la justification *H6-conclusion* (règle 6.d). Ceci est dû au fait que la conclusion comporte typiquement des thèmes qui lui sont propres, comme par exemple la description de travaux ou d'améliorations futurs.

Règle 6.d (Conclusion)

$$(\text{conclusion}(\sigma_i) \wedge \text{thème}(\sigma_i, \varsigma)\{\delta\} \wedge \text{part}(\sigma, \sigma_i)) \quad \Rightarrow \quad \text{thème}(\sigma, \varsigma)\{H6\text{-}conclusion\}.$$

La règle inverse est aussi valable: tous les thèmes d'une section-mère sont aussi thèmes de ses sections-filles (règle 6.e). Ceci permet de propager aux sections-filles les thèmes qui sont propres à la section-mère: par exemple, le titre d'une section, ou les paragraphes de texte qui apparaissent avant la première section-fille.

Règle 6.e (Héritage descendant)

$$\text{thème}(\sigma, \varsigma)\{\delta\} \wedge \text{domine}(\sigma, \sigma_i) \quad \Rightarrow \quad \text{thème}(\sigma_i, \varsigma)\{H6\}.$$

4.3 Le méta-discours dans les collections-tests

Nous examinons dans cette section la fréquence des expressions de méta-discours dans les collections-tests, ainsi que leur incidence sur les requêtes. Pour notre étude nous nous limitons aux expressions formées à l'aide du verbe « *to describe* » (*décrire*) et dont la *localisation* désigne la totalité du document. Les expressions du méta-discours considérées sont donc données par la grammaire:

`localisation` ::= `dét dés`
`dét` ::= ['in'] 'this' | 'the present'
`dés` ::= 'paper' | 'report' | 'note' | 'book' | 'article'
`verbe-descriptif` ::= to describe

4.3.1 Fréquence dans les collections-tests

La table 4.1 donne le nombre d'apparitions de ces expressions pour les collections CACM et *Cranfield*. Chacune des lignes de ce tableau correspond à une forme possible, où l désigne la *localisation*, et i, l'intention.

La colonne «*occ. total*» donne le nombre total de documents qui contiennent l'expression en question. Certains de ces documents contiennent des intentions qui ne correspondent pas au thème principal du document, ou qui ne peuvent être comprises sans une analyse syntaxico-sémantique, afin de résoudre les ambiguïtés et les références anaphoriques. Par exemple le document n° 1046 contient la phrase «*The present paper describes some of the major features of their system*» (*le présent papier décrit quelques-unes des principales caractéristiques de leur système*). Or l'expression «*quelques-unes des principales caractéristiques de leur système*» ne peut pas être utilisée comme telle comme thème, puisqu'il faut savoir à quoi réfère «*leur système*».

Table 4.1. Occurrences d'expressions de méta-discours

forme	*cacm*		*cranfield*	
	occ. total	occ. thème	occ. total	occ. thème
l describes i	63	49	17	17
in l i is/are described	10	4	1	0
i is described in l	7	5	2	2
in l we describe i	1	1	0	0

Afin d'évaluer l'erreur commise en sélectionnant toutes les intentions, nous avons manuellement déterminé parmi la liste quelles intentions pouvaient être utilisées directement comme thème principal du document. Ce résultat apparaît à la colonne «*occ. thème*».

Les expressions de méta-discours apparaissent à 81 reprises dans le corpus de la CACM (somme de la colonne «*occ. total*»); de ce nombre, seules 59 occurrences correspondent au thème principal (colonne «*occ. thème*»), soit une marge d'«*erreur*» assez importante de 27%. Pour la collection Cranfield, ces chiffres sont de 20 et de 19, respectivement, pour une marge d'erreur de 5%. Cette marge d'erreur peut être considérablement diminuée par un processus simple de sélection automatique où les intentions contenant des pronoms personnels ou autres indicateurs de références anaphoriques sont rejetées.

On peut donc obtenir relativement aisément un processus d'extraction qui identifie le thème principal à partir d'expressions du méta-discours, pour 3.7% des documents de la

CACM et 1.4% des documents de la collection Cranfield[5]. Les résultats inférieurs pour la collection Cranfield sont dus à son vocabulaire plus diversifié; pour obtenir de meilleurs résultats il faudrait élargir les expressions de méta-discours.

Rappelons que ces pourcentages ne sont pas des mesures de *précision/rappel*. Une amélioration de précision/rappel de 3.7% par rapport à une ou des requêtes ne serait pas jugée significative, puisqu'elle ne s'appliquerait qu'à l'ensemble des réponses pour ces requêtes. Par contre, une amélioration de 3.7% lors de l'indexation est significative, puisqu'elle s'applique à l'ensemble du corpus.

Si cette analyse préliminaire peut sembler encourageante à prime abord, et ce d'autant plus qu'elle ne considère que des formes très restreintes de méta-discours, il convient d'ajouter un bémol. Les deux collections étudiées ici ne contiennent que des résumés d'articles: non seulement est-il beaucoup plus probable d'y trouver des expressions de méta-discours que dans du texte intégral, mais de plus, les expressions y sont généralement plus simples et plus facilement identifiables.

4.3.2 Incidence du méta-discours sur les requêtes

Nous voyons maintenant l'impact que peuvent avoir les expressions de méta-discours sur les requêtes. Nous proposons une stratégie de recherche où les thèmes provenant des expressions de méta-discours sont favorisés. Pour ce faire, en plus de l'indexation des documents entiers, une seconde indexation est réalisée sur les intentions données par les expressions de méta-discours. La recherche s'effectue indépendamment sur les deux fichiers d'indexation, puis les deux listes de documents ainsi obtenues sont combinées en plaçant en tête de liste les documents issus des intentions.

Cette stratégie est comparée à l'approche traditionnelle à la figure 4.2 pour la requête n° 23 de la CACM, «*Distributed computing structures and algorithms*» (*Structures et algorithmes informatiques distribués*). Cette requête est traduite par «<0.7> #AND (distributed computing structures algorithms)» dans notre système.[6]

La figure montre la *précision*, c'est-à-dire la proportion de documents pertinents parmi les documents retrouvés, en fonction du *rappel*, ou le nombre de documents pertinents retrouvés sur le nombre de documents pertinents total, qui sont au nombre de 4 pour la requête n° 23. Notre stratégie est montrée par la ligne continue, alors que l'approche traditionnelle est montrée par la ligne grisée. Pour le premier document pertinent retrouvé – où le rappel est de 0.25 – on obtient une précision de 0.1 par notre approche, et de seulement 0.0008 par l'approche traditionnelle. Ceci est dû à la différence de classement du document n° 3148: celui-ci n'est classé que très loin dans la liste selon l'approche traditionnelle, mais

5. Soit 59 documents sur 1587 pour la collection CACM – puisque des 3204 documents, seuls 1587 comportent un résumé – et 19 documents sur 1398 pour Cranfield.

6. Le poids sur l'opérateur AND permet une sélection à cheval entre la conjonction – trop stricte – et la disjonction – trop large. Pour plus de détails voir l'annexe B.

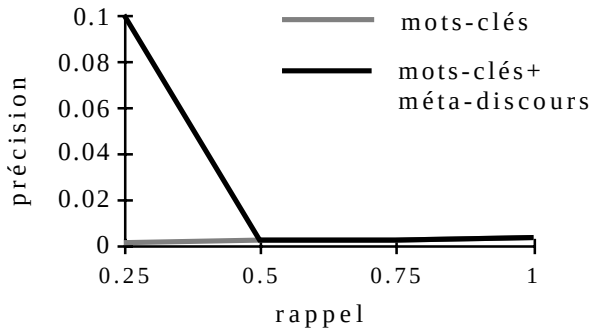


Figure 4.2. Précision/rappel pour la requête n°23 de la CACM

il est promu au 10e rang selon notre approche, parce qu'il contient l'expression « *this paper describes an approach to distributed computing at the level of general purpose programming languages* » (*ce papier décrit une approche d'informatique distribuée au niveau des langages de programmation à usage général*). Par la suite, les deux courbes montrent la même précision.

Malheureusement cette stratégie a peu d'impact globalement sur l'ensemble des requêtes de la CACM. Ceci s'explique par le fait que seules 4 des 52 requêtes ont dans leur liste de documents pertinents un document contenant une expression de méta-discours (soit une proportion plus élevée que le pourcentage du corpus qui contient de telles expressions). La figure 4.3 montre la courbe combinée de précision/rappel pour ces quatre requêtes, soit les n^{os} 7, 12, 16, et 23.

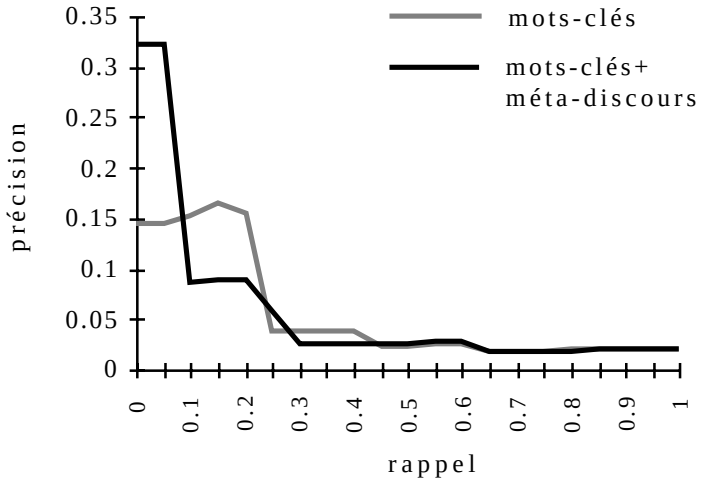


Figure 4.3. Précision/rappel pour les requêtes n^{os} 7, 12, 16, et 23 de la CACM

On peut constater que la précision est généralement meilleure par notre approche pour les premiers documents: ceci est vérifié sur trois des quatre requêtes. Par la suite, les deux courbes se suivent, la précision de par notre approche pouvant même être légèrement inférieure. Notre approche est donc avantageuse lorsqu'il est important pour l'utilisateur de retrouver au moins un document pertinent en tête de liste.

4.4 Validation des règles auprès d'utilisateurs

Nous discutons dans cette section d'une expérience ayant pour but de valider les règles définies dans ce chapitre, c'est-à-dire de vérifier si elles correspondent bien à la sélection de thèmes telle que faite par les utilisateurs. Dans cette expérience, des volontaires devaient identifier le ou les thèmes dans le texte: soit en sélectionnant parmi les choix proposés le meilleur thème pour une expression, soit en dégageant eux-mêmes les thèmes d'un court texte.

Puisque nous avons défini des règles spécifiques au français et à l'anglais, nous avons cru intéressant de réaliser deux versions pour l'expérience; l'une en français et l'autre en anglais. Il convient de spécifier toutefois que ces deux expériences doivent êtres considérées indépendamment l'une de l'autre: les deux groupes de volontaires étant très différents, et les spécificités de chaque langue pouvant jouer. Nous décrivons ci-dessous les deux expériences, mais en mettant l'emphase sur l'expérience en français. Pour les détails spécifiques à l'expérience en anglais on peut consulter [PB96].

4.4.1 Méthodologie de l'expérience

Une des particularités de notre expérience est que le questionnaire a été implémenté par un formulaire SGML et rendu accessible par le biais du World Wide Web. Nous pensons que cette méthode peut attirer une plus grande variété de participants, en leur fournissant la facilité et le confort de leur propre environnement pour répondre au questionnaire. Le désavantage est le manque de contrôle sur les participants et leur environnement; ils peuvent par exemple être interrompus pendant l'expérience, ou faire autre chose en parallèle.

Un autre danger de cette méthodologie est d'attirer les *net surfers*, c'est-à-dire des gens qui n'ont pas été sollicités pour l'expérience, mais qui trouvent l'adresse électronique par hasard, et qui pourraient remplir le questionnaire sans le sérieux qu'il mérite. Ainsi l'adresse électronique n'a-t-elle été communiquée qu'à un nombre restreint de personnes. De plus les questionnaires incomplets étaient automatiquement refusés; et dans un cas où le participant semblait avoir des réponses incohérentes, le questionnaire a été rejeté.

Plusieurs participants ont signalé avoir eu des problèmes avec leur programme de navigation, et près d'une dizaine de questionnaires ont été perdus de cette façon. En fait, dû à l'utilisation de tableaux imbriqués dans notre formulaire, et à une erreur de *Netscape* sur

Mac, ou de l'utilisation de navigateurs obsolètes comme *Mosaic*, le questionnaire dans sa version initiale n'était pas utilisable par tous. Une version équivalente mais dont l'impact visuel était moindre a été réalisée pour remédier à cette situation.

Afin de recruter les participants, un message de sollicitation a été envoyé sur diverses listes de distribution. On pourrait soutenir qu'il n'était pas approprié de demander une tâche pareille à des non-linguistes. Bien que ceci serait vrai si le but avait été l'analyse de texte, nous ne croyons pas que ce soit le cas pour la recherche d'informations. Ce sont souvent des non-linguistes et des non-spécialistes qui décident finalement si les documents retournés sont pertinents ou pas. Il est donc aussi important de modéliser leur jugement de pertinence que celui des experts.

L'expérience en anglais a eu une durée d'un mois. Le message de sollicitation a été envoyé aux groupes suivants:

- les participants à l'école d'été européenne de recherche d'informations (ESSIR), tenue à Glasgow en septembre 1995;
- les participants au colloque MIRO, aussi tenu à Glasgow immédiatement après ESSIR;
- les participants à la conférence LLI (Logic, Language and Information), tenue à Espinho, Portugal en décembre 1994.

Les deux premiers groupes visaient des gens avec des intérêts en recherche d'informations, alors que le troisième groupe était surtout formé de logiciens et de linguistes. En tout, ces trois groupes comprenaient 178 personnes, et de ce nombre, 42 ont accepté d'être volontaires pour l'expérience.

Pour l'expérience en français, le recrutement s'est largement fait parmi la population étudiante de l'IMAG et de l'université de Montréal. 34 personnes ont répondu au questionnaire, sur une durée de plus de 8 mois.

Le questionnaire consiste en deux parties: une première partie comporte 20 questions où les participants doivent choisir un thème parmi le choix qui leur est proposé, et une seconde où ils doivent eux-mêmes dégager les thèmes d'un court texte. Le questionnaire français est une traduction – la plus fidèle possible – du questionnaire anglais. Les deux formulaires sont donnés à l'annexe A. Les répondants ont mis en moyenne 20.4 minutes pour compléter le questionnaire (19.2 pour les anglophones et 21.9 pour les francophones).

4.4.2 Distribution des participants

Avant de répondre au questionnaire comme tel, les participants devaient d'abord s'identifier, en donnant leur nom et coordonnées, et en répondant à quelques questions ayant pour but de déterminer leurs compétences vis-à-vis de l'expérience.

Les 42 participants de l'expérience en anglais étaient originaires de 11 pays différents distribués comme suit: 66% d'Europe, 24% des Amériques, et 10% d'Asie. La majorité des répondants francophones (85%) provenaient de France.[7]

La première question posée aux participants visait à déterminer leur maîtrise de la langue. La question posée était la suivante: «*Comment évaluez-vous votre maîtrise du français (de l'anglais) écrit?*». Les résultats pour les deux groupes sondés sont présentés à la figure 4.4. Les proportions sont à peu de choses près les mêmes pour les groupes français et anglais: 2/3 des répondants disent avoir une «*excellente*» maîtrise de la langue, 1/3 en ont une «*bonne*» maîtrise, et une faible proportion une maîtrise «*limitée*». Cette répartition doit être interprétée différemment pour le français et pour l'anglais. Alors que pour l'anglais elle distingue les *natifs* de ceux dont l'anglais n'est pas la langue première, pour le français, elle ne reflète qu'un degré de compétence entre francophones.

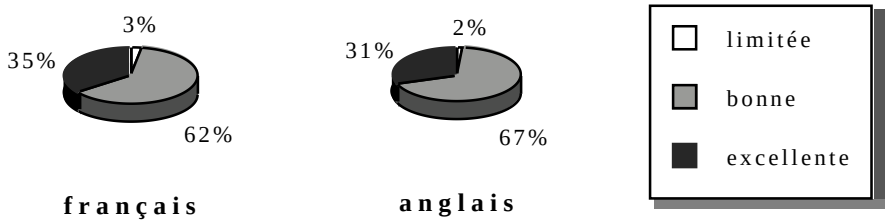


Figure 4.4. Maîtrise de la langue écrite

Il s'avère que les résultats de l'expérience ne semblent pas varier selon la compétence linguistique, à l'exception des répondants dont la maîtrise de la langue est «*limitée*». Heureusement ces participants ne représentent pas une proportion significative des répondants.

La seconde question portait sur la familiarité avec la recherche d'informations. Les résultats sont présentés à la figure 4.5. Trois choix étaient possibles, allant d'une connaissance nulle à une connaissance très avancée de la recherche d'informations. Pour l'expérience en français, les répondants sont à peu près répartis entre les trois choix de réponses; alors que pour l'expérience en anglais les répondants se répartissent sur les deux derniers choix, avec seulement 7% de répondants pour le choix «*pas du tout*». Ceci n'est pas surprenant étant donné les personnes rejointes par notre sollicitation.

La troisième et dernière question portait sur la familiarité avec les outils de recherche documentaire. Les résultats sont présentés à la figure 4.6. La question était: «*Avez-vous déjà utilisé un système de recherche documentaire, dans une bibliothèque ou sur le Web?*». Il n'est pas surprenant de constater qu'une faible proportion des répondants dit n'avoir jamais utilisé de tel système, puisque, devant utiliser le Web pour répondre au questionnaire, ils ont toutes les chances d'être familiers avec ces outils.

7. L'origine est établie uniquement d'après l'adresse électronique.

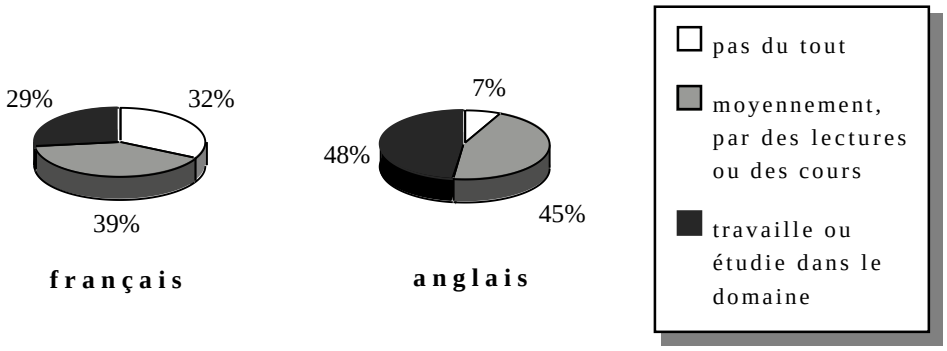


Figure 4.5. Familiarité avec la recherche d'informations

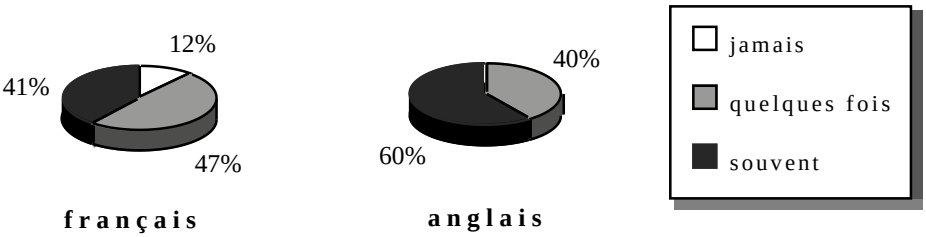


Figure 4.6. Familiarité avec les systèmes de recherche d'informations

4.4.3 Les thèmes dans des expressions

Dans la première partie du questionnaire, les participants devaient choisir lequel de deux items proposés représentait le mieux le thème ou le sous-thème d'une expression donnée. Les expressions étaient choisies pour donner le contexte approprié; elles consistaient en une ou deux phrases.

Par exemple, soient les choix «*voiture*» et «*rouge*» pour l'expression «*une voiture rouge*». Si l'hypothèse de dépendance (H2) est bien fondée, alors on devrait préférer «*voiture*» à «*rouge*», puisque l'adjectif «*rouge*» est dépendant du nom «*voiture*» dans cette expression.

Un troisième choix, «*également*», est utilisé lorsque les deux items proposés ne peuvent pas être discriminés. Ainsi, dans la question n° 10 du questionnaire,

Q10. Les trous noirs sont formés par l'effondrement d'étoiles massives.

- ◯ noirs
- ◯ trous
- ◯ également
- ◯ ne sais pas

il n'est ni question de «*noir*», ni de «*trous*», mais bien d'un corps céleste, «*trou noir*». On voit dans cet exemple un quatrième choix, «*ne sais pas*», que le répondant utilise lorsqu'il ne peut pas répondre à la question.

Le questionnaire comporte 20 de ces questions, qui permettent de valider les règles de contenu sémantique (H1), de dépendance (H2), d'analyse statutaire (H3) et de progression thématique (H4). Ces règles sont distribuées aléatoirement parmi les 20 questions.

a) Distribution générale des réponses

La figure 4.7 présente les réponses aux 20 questions pour les expériences en français et en anglais. Chacune des 20 questions comporte quatre bâtons, correspondants aux quatre choix de réponses. Bien entendu les «*choix 1*» et «*choix 2*» diffèrent d'une question à l'autre. Ainsi, pour la question n° 10, ces choix correspondent respectivement à «*noirs*» et à «*trous*». Voici la distribution des réponses pour cette question pour l'expérience en français: 0% ont choisi «*noirs*», 9% ont choisi «*trous*», 85% ont choisi «*également*», et 6% n'ont pas pu répondre. Les résultats sont sensiblement les mêmes pour cette question en anglais.

Les questions françaises sont généralement des traductions de leur contrepartie anglaise, sauf pour les questions n^{os} 8 et 12, qui dans la version anglaise font référence à «*blind Venetians*» (*vénitiens aveugles*, un jeu de mot par rapport à «*venetian blind*», *store vénitien*), et pour la question n° 19, où l'ambiguïté entre *tasse* et *café* n'est pas traduisible en français à cause de l'accord en genre.

Un rapide coup d'œil aux graphes de la figure 4.7 révèle que le choix «*ne sais pas*» n'a pas récolté beaucoup de réponses, fort heureusement d'ailleurs puisqu'un fort taux de ces réponses invaliderait notre questionnaire. En moyenne les réponses «*ne sais pas*» représentent 3% de toutes les réponses pour le français, et 6% pour l'anglais. Cependant, plusieurs participants ont remarqué qu'ils ne pouvaient pas bien distinguer entre «*également*» et «*ne sais pas*», aussi le taux de réponses incertaines est peut-être en réalité légèrement supérieur.

La courbe superposée aux histogrammes à la figure 4.7 trace le nombre de répondants pour le choix *prédominant*, c'est-à-dire le choix 1 ou 2, selon le cas, après redistribution des réponses «*également*» et «*ne sais pas*». En moyenne, le choix prédominant récolte 69% des répondants pour le français, et 67% pour l'anglais, ce qui montre une tendance significative pour les répondants à faire le même choix. Les deux exceptions sont les questions n^{os} 5 et

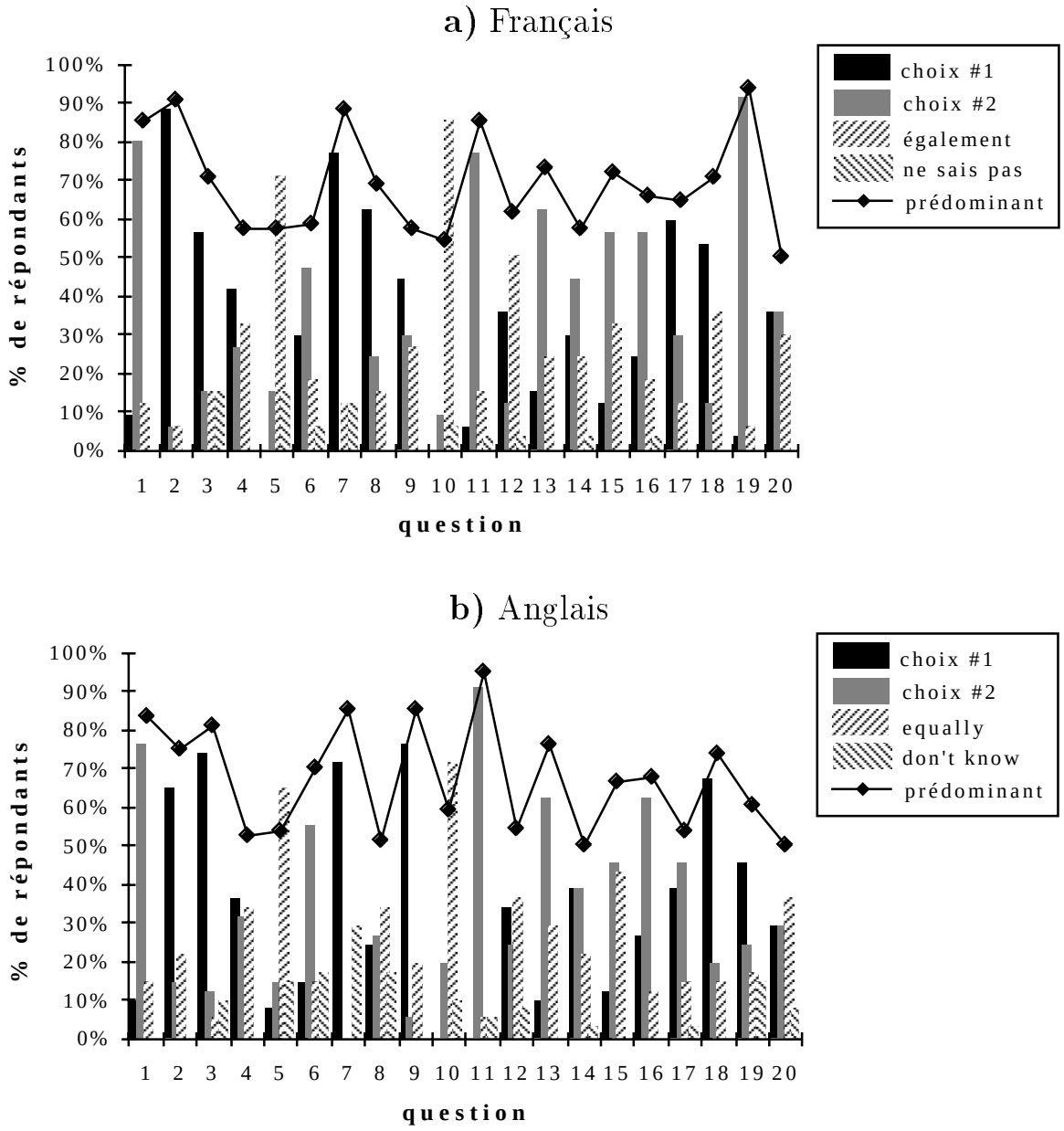


Figure 4.7. Réponses pour la partie I du questionnaire

10, où la réponse espérée était «*également*» (voir la section *Autres aspects* plus bas).

Nous voyons maintenant comment les règles ont été validées dans l'expérience.

b) Règle de contenu sémantique

La règle de *contenu sémantique* (H1) était testée aux questions n^{os} 3, 6, 7 et 17. Les répondants semblent utiliser le contenu sémantique et les connaissances de façon assez cohérente. Lorsqu'ils ont le choix entre des synonymes et des mots liés ils choisissent généralement le même terme. Ils peuvent même choisir un terme lié s'il leur semble plus significatif que le terme original sorti de son contexte.

À la question n° 3 ci-dessous, les répondants préfèrent «*Bordeaux*» à «*Europe*», ce qui suggère que la relation entre «*Bordeaux*» et «*vins*» ou entre «*Bordeaux*» et «*France*» est plus forte que la généralisation *France* $\supset$ *Europe*.

Q3. La France est bien connue pour ses vins.

- ◯ Bordeaux
- ◯ Europe
- ◯ également
- ◯ ne sais pas

Ceci semble suggérer que les inférences logiques du genre *spécialisation/généralisation*, qui sont souvent présumées valides dans les modèles de recherche d'informations, peuvent être éclipsées par d'autres facteurs. La question n° 17 supporte partiellement cette idée. Ici le choix est à faire entre «*baleines*» et «*baleines bleues*», où chacun de ces termes apparait comme thème de la phrase. Les répondants français choisissent «*baleines*» à 59%, mais les répondants anglais ne peuvent pas décider pour l'un ou l'autre (démontrant même une légère préférence pour «*baleines bleues*»), malgré le fait que «*baleines*» soit une généralisation de «*baleines bleues*». Chez les répondants possédant une maîtrise de la langue *excellente*, le choix est plus marqué sur «*baleines*» (75% pour le français, et 54% pour l'anglais).

Pour les questions n^{os} 6 et 7, les répondants préfèrent un choix qui n'apparaissait pas dans l'expression initiale. Ainsi à la question n° 6, pour l'expression «*Mont Everest*», les répondants préfèrent «*montagne*» à «*Mont*». À la question n° 7, toujours pour «*Mont Everest*», il y a une préférence marquée pour «*alpinisme*» sur «*ski*».

c) Règle de dépendance

La règle de *dépendance* (H2) est testée aux questions n^{os} 1, 2, 8, 11, 15 et 19. Les résultats supportent généralement cette idée de dépendance, pour autant qu'elle soit appliquée de façon *intelligente*, c'est-à-dire pas seulement d'après des règles syntaxiques ou grammaticales figées, mais aussi en considérant la sémantique, voire même la pragmatique.

La question n° 15 supporte l'hypothèse de dépendance; De l'expression « *vins de France*», où «*France*» est dépendant de «*vins*», les participants préfèrent «*vins*». À la question n° 8 ci-dessous, la réponse est moins simple: s'agit-il d'un «*aveugle Vénitien*» ou d'un «*Vénitien aveugle*»?[8]

Q8. Blind Venetians are a rare sight at Mount Everest.

- ◯ blind
- ◯ Venetians
- ◯ également
- ◯ ne sais pas

La plupart des répondants n'ont pas su discriminer les deux choix: 24% ont choisi «*blind*», 26% «*Venetians*», le reste étant partagé entre «*également*» et «*ne sais pas*». Pour les anglais «*natifs*» cependant, ces résultats sont de 8% et 23%, respectivement.

Pour certaines expressions le choix peut dépendre du contexte. C'est le cas par exemple de «*tasse de café*», qui apparaît aux questions n^{os} 1, 11 et 19 (pour le questionnaire anglais). À la question n° 1 ci-dessous, «*tasse*» est préférée par une forte majorité des participants. Les mêmes résultats sont aussi observés à la question n° 19.

Q1. Une tasse de café a été laissée sur la table.

- ◯ café
- ◯ tasse
- ◯ également
- ◯ ne sais pas

Par contre, pour la question n° 11, cette fois «*café*» est préféré avec une majorité tout aussi marquée.

Q11. Le repas peut être suivi d'une tasse de café.

- ◯ tasse
- ◯ café
- ◯ également
- ◯ ne sais pas

De même pour la question n° 2 ci-dessous, malgré le fait que «*diesel*» soit dépendant de «*voitures*», une grande majorité l'a choisi comme thème.

Q2. Les voitures diesel sont généralement considérées comme étant plus dommageables pour l'environnement.

- ◯ voitures
- ◯ diesel
- ◯ également
- ◯ ne sais pas

Ces exemples montrent que la règle de dépendance varie selon le contexte et donc qu'elle

8. L'anglais n'admet pas cette dernière formulation, puisque «*Venitian blind*» a le sens de *store vénitien*, d'où l'ambiguïté.

peut être difficile à automatiser. On peut cependant remarquer que les deux derniers contre-exemples, les questions n^{os} 11 et 2, pourraient être paraphrasés par «*Le repas peut être suivi de café*» et «*Les diesel sont généralement...*», respectivement, ce qui lève le problème.

d) Règle d'analyse statutaire

L'hypothèse d'*analyse statutaire* est testée aux questions n^{os} 12, 13 et 18. Pour la question n^{o} 18 ci-dessous, «*Paris*» appartient au thème, alors que «*cafés*» appartient au rhème. Les participants ont bien choisi «*Paris*», et ce malgré toute l'information qui était ajoutée à «*cafés*». Des résultats similaires sont obtenus pour la question n^{o} 13; le thème, «*Jupiter*», bien qu'il joue cette fois le rôle de complément, est choisi par 62% des participants.

Q18. Paris possède beaucoup de cafés agréables avec terrasse et à l'atmosphère détendue.

- ◯ Paris
- ◯ cafés
- ◯ également
- ◯ ne sais pas

Pour la question n^{o} 12, les deux items appartiennent au thème: les participants ont eu tendance à choisir le sujet de la phrase, «*Edelweiss*» pour l'expérience français, et «*Blind Venetians*» pour l'expérience anglais.

e) Règle de progression thématique

Les questions n^{os} 4, 9, 14, 16 et 20 testent l'hypothèse de *progression thématique* (H4). Les résultats sont peu concluants: seulement pour la question n^{o} 9, qui était une combinaison *sujet-sujet*, avons-nous obtenu les résultats escomptés: les répondants ont choisi le sujet, «*Pluton*», plutôt que le commentaire, «*Système Solaire*».

Q9. Pluton est la planète la plus éloignée que nous connaissions dans le Système Solaire. C'est aussi la plus petite du Système Solaire.

- ◯ Pluton
- ◯ Système Solaire
- ◯ également
- ◯ ne sais pas

Les résultats pour d'autres combinaisons, *sujet-commentaire* ou *commentaire-sujet*, ne sont pas concluants, c'est-à-dire que les participants ne sont pas arrivés à un concensus sur leur choix. Toutefois, pour la question n^{o} 16, une combinaison *commentaire-commentaire*, le second commentaire, «*ALENA*» semble avoir été privilégié.

Nous avons étudié les corrélations possibles entre les réponses, afin de voir si un répondant donné était cohérent parmi ses réponses. En particulier, il était espéré qu'une

corrélation existerait entre les réponses des questions n^{os} 4, 14, 16 et 20. Malheureusement les différents tests statistiques que nous avons effectués ne nous permettent pas de conclure à cette corrélation. Tout au plus avons-nous trouvé quelques liens marginaux. Ainsi dans l'expérience en anglais, les participants ayant répondu «*dauphins*» à la question n° 4 tendaient à choisir «*baleines bleues*» à la question n° 14; et de même ceux ayant répondu «*pollution de l'eau*» choisissaient «*chasse*». Dans l'expérience en français, les participants ayant répondu «*baleines bleues*» à la question n° 14 choisissaient généralement «*activistes de GreenPeace*» à la question n° 20; et de même ceux qui répondaient «*chasse*» choisissaient «*Muroroa*».

f) Autres aspects

Deux questions, les n^{os} 5 et 10, étaient délibérément choisies de façon à ce qu'aucun des deux items ne puisse être discriminé par rapport à l'autre. Nous avons déjà vu la question n° 10 ci-dessus. Pour la question n° 5, l'expression était «*onze novembre*», où «*onze*» et «*novembre*» ne peuvent ni l'un ni l'autre être considérés comme thème de l'expression.

Nous croyons aussi que l'usage des noms propres dans certaines questions a pu influencer le choix des participants. Dans tous les cas où les participants avaient à choisir entre un nom propre et un nom commun, ils ont choisi le nom commun: que ce soit «*Jupiter*» à la question n° 13, «*ALENA*» à la question n° 16, etc.[9] Ceci peut s'expliquer par l'opposition entre l'information *donnée* versus l'information *nouvelle*): le nom propre faisant toujours référence à un référent supposé connu ou accessible du lecteur, et donc *donné*.

4.4.4 Les thèmes dans un texte

Dans la seconde partie du questionnaire, un court texte humoristique à propos du Père Noël était présenté et les participants devaient identifier les thèmes pour ce texte. Le texte consistait en un paragraphe introductif suivi de deux sous-sections. Les participants devaient donner les thèmes séparément pour le texte entier et pour les deux sous-sections; il leur était de plus demandé de distinguer les thèmes *principaux* des thèmes *secondaires*. Aucun indice n'était donné sur la façon d'utiliser la structure logique pour dériver les thèmes.

Les règles d'intention (H5) et de structure (H6) sont testées ici, en plus de quelques autres données intéressantes comme la cohérence entre les participants.

a) Distribution générale des thèmes

Les thèmes identifiés par les répondants vont du simple mot-clé comme «*rennes*» aux phrases complètes comme «*il fait froid au Pôle Nord*», en passant par le style télégra-

9. La seule exception est «*France*» à la question n° 15 (dans l'expression «*vins de France*»).

phique comme dans «*santa = north pole? ha!*». Certaines réponses avaient plus à voir avec la caractérisation ou le genre du texte qu'avec ses thèmes: par exemple «*humour*», «*anthropologie*», «*pseudo-science*», etc. Les principaux thèmes utilisés en anglais et en français sont équivalents: dans les deux cas, ce sont «*Père Noël*», «*Pôle Nord*», «*rennes*», «*Laponie*», etc.

Un survol de ces thèmes montre que les participants ont beaucoup *paraphrasé*. Dès qu'ils cherchent à exprimer une idée avec plus de 2 mots, plutôt que d'utiliser les mots du texte, ils reformulent dans leurs propres mots. Certains thèmes ne sont que des sous-thèmes d'autres thèmes; par exemple «*vitesse*» et «*vitesse de distribution*».

La table 4.2 donne le nombre d'occurrences des thèmes parmi les réponses des participants pour le français et l'anglais. Les thèmes sont répartis selon le passage d'où ils sont tirés – *texte entier*, *section 1*, *section 2* – et leur importance – *principaux* ou *secondaires*. La colonne *total* est obtenue en combinant les thèmes principaux et secondaires pour un passage. La ligne **total**, elle, correspond à la combinaison des thèmes pour chacun des trois passages.

Deux occurrences sont données: le nombre de thèmes *distincts* et, entre parenthèses, le nombre de thèmes *total*. Les termes *distincts* sont obtenus après avoir «*normalisé*» certains articles ou prépositions. Ainsi, «*les cadeaux du Père Noël*», «*cadeaux du Père Noël*», et «*cadeaux de Père Noël*» ne comptent que pour 1 thème distinct (mais pour 3 thèmes total). Les paraphrases comptent toujours pour des thèmes distincts: par exemple, «*nuit polaire*» et «*nuit au Pôle Nord*».

Table 4.2. Nombres de thèmes distincts (total)

		principaux	*secondaires*	*total*
français	*texte entier*	22 (50)	67 (83)	83 (133)
	section 1	31 (54)	54 (104)	77 (158)
	section 2	36 (50)	96 (128)	123 (178)
	total	81 (154)	195 (315)	248 (469)
anglais	*texte entier*	22 (51)	65 (100)	81 (151)
	section 1	27 (49)	60 (124)	81 (173)
	section 2	47 (53)	101 (133)	141 (186)
	total	86 (153)	211 (357)	279 (510)

Du texte original français, qui comptait 584 mots, les participants ont extrait 248 thèmes distincts. Pour ce qui est du texte anglais, qui avait une longueur de 565 mots, 279 thèmes distincts ont été identifiés. En moyenne les participants ont identifié 1.5 thèmes principaux

en français, et 1.2 en anglais, contre 3.1 thèmes secondaires en français, et 2.8 en anglais. Ceci semble indiquer que la consigne de différencier thèmes principaux et secondaires a été bien suivie.

La table 4.3 donne l'occurrence moyenne et l'écart-type des thèmes distincts. Au total, pour l'ensemble des thèmes français, l'occurrence moyenne était de 1.9 et l'écart-type de 3.2; ces chiffres sont semblables (1.8 et 3.1) pour l'expérience en anglais. Ceci signifie qu'en moyenne les thèmes étaient uniques (moyenne près de 1) sauf pour quelques thèmes utilisés par beaucoup de répondants (écart-type élevé).

Table 4.3. Occurrence moyenne (écart-type) des thèmes distincts

		principaux	*secondaires*	*total*
français	*texte entier*	2.3 (3.8)	1.2 (0.8)	1.6 (2.3)
	section 1	1.7 (1.7)	1.9 (2.9)	2.1 (2.7)
	section 2	1.4 (1.4)	1.3 (0.9)	1.4 (1.2)
	total	1.9 (3.7)	1.6 (2.0)	1.9 (3.2)
anglais	*texte entier*	2.3 (5.2)	1.5 (1.5)	1.9 (3.1)
	section 1	1.8 (1.7)	2.1 (3.3)	2.1 (3.4)
	section 2	1.1 (0.3)	1.3 (1.0)	1.3 (1.0)
	total	1.8 (3.6)	1.7 (2.3)	1.8 (3.1)

b) Règle d'intentions

Pour tester la règle d'*intentions* (H5), nous avons vérifié l'utilisation de certains expressions provenant des titres ou du méta-discours. Ces expressions n'étaient utilisées qu'à un seul endroit dans le texte, aussi si les répondants les ont utilisées, il y a une forte chance qu'elles proviennent de cet endroit.

Le titre principal était «*Le Père Noël: Mythes & Réalités*». Nous avons recensé l'usage des trois termes de ce titre: «*Père Noël*», «*mythes*» et «*réalités*». Pour l'expérience en français, seuls 12% des répondants sélectionnent les trois termes comme thèmes du texte entier, 35% en sélectionnent deux, 41% un seul, et 12% aucun. Pour l'anglais, ces chiffres sont respectivement de 10%, 40%, 50% et 0%.

Le titre de la section 1 était «*L'emplacement du Père Noël*»: il a été utilisé par 21% des répondants (29% pour l'anglais). Le titre de la section 2, «*Comment fait-il?*», a délibérément été choisi comme non-significatif, il s'avère toutefois que trois participants (1 français et 2 anglais) l'ont identifié comme thème.

Les résultats quant à l'emploi d'expressions de méta-discours sont moins concluants. Trois expressions ont été placées dans le texte comme des indications directes de l'*intention* de l'auteur, mais peu de répondants les ont utilisées.

Au premier paragraphe, la phrase «*Nous discutons dans ce court texte de deux conceptions erronées concernant le Père Noël*» donne le but général du texte. 12% des répondants français ont identifié «*conceptions erronées*» comme thème principal, et 11% des répondants anglais ont fait de même pour «*misconceptions about Santa Claus*». La dernière phrase de la section 1 est «*Ayant discuté du véritable emplacement du Père Noël, nous allons maintenant examiner dans la prochaine section le "problème de distribution des cadeaux"*»; elle contient à la fois un sommaire de la section 1, et une présentation de la section 2. Il serait hasardeux de tirer des conclusions de l'emploi du thème «*emplacement du Père Noël*» en français, puisque cette expression apparaissait également comme titre de la section. En anglais, où le titre de la section était «*The location of Santa Claus*», et la phrase finale «*Having discussed the real location of Santa's home...*», 5% ont choisi «*location of Santa's home*» comme thème. Quand au second indice, pour la section 2, il a été utilisé par 24% en français, et seulement 2% en anglais.

c) Règle de structure

Nous cherchons ici à valider la règle de *structure* (H6). Dans les instructions pour remplir le formulaire, aucune indication n'est donnée aux participants sur la façon dont ils peuvent déduire les thèmes pour le texte entier: ils peuvent donc le faire uniquement en se basant sur le titre principal et le paragraphe introductif, ou en utilisant également les deux sections.

Nous avons vérifié si les répondants utilisaient les thèmes principaux des sections pour dériver les thèmes secondaires du texte entier. 15% des répondants ont réutilisé tous les thèmes principaux comme thèmes secondaires du texte, 24% en ont réutilisé quelques-uns, et 62% pas du tout. Pour l'anglais, ces nombres sont de 17%, 29% et 54%.

d) Apparence physique du texte

Enfin, pour tester comment les lecteurs peuvent être influencés par l'apparence ou l'arrangement du texte, nous avons mis l'expression «*problème du voyageur de commerce*» en italique à la section 2. Cette expression n'était pas pertinente pour le texte – il était dit en fait que l'on *ignorait* le problème de planification de l'itinéraire – pourtant 18% des répondants l'ont choisie comme thème (24% pour l'anglais).

4.4.5 Conclusion sur la validation

Notre validation montre que bien qu'il y ait de grandes variations entre les lecteurs, la plupart d'entre eux suivent le même raisonnement pour la dérivation des thèmes. Nos règles de dérivation semblent de bonnes approximations de ce raisonnement, puisqu'elles arrivent généralement aux mêmes conclusions que les lecteurs.

Une exception toutefois est la règle de *progression thématique*; les résultats ne nous permettent pas de conclure à sa validité – mais ils ne la réfutent pas non plus. Puisque les questions portant sur cette règle étaient forcément plus longues – elles comportaient typiquement deux phrases – d'autres facteurs peuvent avoir joué ici; il se peut également qu'elle ne soit pas assez *fine*.

Même s'il y avait concensus sur les réponses parmi les lecteurs, souvent une minorité significative d'entre eux préférait une autre réponse. Nous attribuons principalement cet écart aux différents contextes dans lequel se plaçaient les utilisateurs face à l'énoncé qu'ils lisaient.

4.5 Conclusion

Dans ce chapitre, nous avons défini les règles de dérivation de notre modèle. Si ces règles s'apparentent à des systèmes de production, puisqu'à partir de faits connus elles permettent de dériver de nouveaux faits, à cause de leur simplicité, elles en éludent les principaux inconvénients [WB88]. Le problème de la *cohérence* ne se pose pas dans notre modèle, puisqu'il n'est pas possible de dériver de fait négatif. Quant au problème de *priorité* ou de *critère d'arrêt* dans l'application des règles, il est évité du fait que les règles ne soient pas récursives.

Nos règles ne doivent pas non plus être confondues avec les algèbres de bases de données textuelles (*text databases algebra*) [ST92, CCB94]. Ces dernières sont plutôt concernées par l'ordre d'apparition et les relations structurelles qui existent entre les mots, tandis que nos règles visent l'extraction des idées prédominantes d'un document. D'autres travaux plus similaires concernent la recherche de passages (*passage retrieval*) ou de texte intégral (*full text retrieval*) [Cal94, Wil94, SAB93]: ces travaux démontrent aussi l'utilité de la prise en compte de la structure dans la dérivation des thèmes.

Nous nous sommes concentrés ici exclusivement sur la dérivation d'index *thématiques*. Cette contrainte ne nous paraît pas trop restrictive puisque cette notion est si importante que la recherche d'informations est en fait souvent réduite à une recherche thématique. Il serait de plus impossible dans l'état actuel des connaissances d'élaborer un modèle d'indexation général pour tout type de requête, puisque certains types d'index nécessitent une représentation sémantique plus poussée, dépendante du domaine d'application. Enfin, notons que cette restriction n'enlève rien à la généralité du langage de représentation $\mathcal{L}$ présenté au chapitre précédent; ce dernier permet la représentation de tout index et de

toute information permettant de dériver cet index.

L'application des règles de méta-discours à la collection CACM, semble indiquer une augmentation de la précision par notre approche. Notre validation avec les utilisateurs, quant à elle, suggère le besoin de modéliser des règles *dynamiques* d'indexation, c'est-à-dire qui s'adaptent à l'utilisateur et au contexte de sa requête. Nos règles forment un bon cadre pour permettre un premier pas vers cette modélisation.

Chapitre 5

Application à un corpus technique

L'expérience est le nom que chacun
donne à ses erreurs.

Oscar WILDE
(*L'éventail de Lady Widermere*)

Le but de ce chapitre est de montrer expérimentalement la validité de notre modèle. Puisque le processus d'indexation n'est qu'une étape intermédiaire et transparente à l'utilisateur, il est difficile de l'évaluer indépendamment d'un système complet. Nous proposons donc ici un prototype intégrant les fonctions d'indexation, de correspondance, de même qu'une interface de formulation de requête et de visualisation des réponses.

Une autre difficulté relève du choix de la collection. Le problème de la disponibilité des documents électroniques structurés ne se pose pas vraiment, étant donné que des formats largement répandus comme LaTeX ou HTML intègrent dans une certaine mesure les informations utilisées par notre modèle. Par contre, il n'existe toujours pas à ce jour de collection standard[1] qui tienne compte de ces informations. Notre approche ne peut donc être appliquée dans l'immédiat qu'à un ensemble de textes pour lequel on ne dispose pas de requêtes et de jugements de pertinence. Ceci nous oblige à baser notre évaluation sur des requêtes que nous avons nous-mêmes définies; cette évaluation est réalisée dans une optique de comparaison avec d'autres approches et non pour l'obtention de résultats absolus.

Nous donnons d'abord les grandes lignes de notre expérimentation, en présentant la collection choisie ainsi que l'architecture générale de notre prototype, puis détaillons les processus de représentation des documents et d'évaluation des requêtes. Enfin nous analysons les résultats, en comparant notre approche à un système de recherche textuelle basé sur PAT [Loe94].

1. Nous entendons par collection standard un ensemble homogène de documents pour lequel des requêtes et leur jugement de pertinence associé ont été définis.

5.1 Principes généraux

5.1.1 Présentation de la collection

Nous avons choisi d'appliquer notre prototype à un manuel technique, les «*recommandations pour l'échange et la représentation de textes électroniques*» (*TEI Guidelines*) , auquel nous référons par les «*Recommandations TEI*» ci-dessous.[2] Ce manuel consiste en 33 chapitres décrivant le formalisme TEI pour l'encodage et l'échange de documents électroniques, et est lui-même encodé dans le formalisme TEI. La taille des chapitres varie de moins de 5 Ko à près de 250 Ko. En tout la collection fait plus de 2.23 Mo, dont à peu près 6% pour le marquage, c'est-à-dire que le contenu purement textuel des chapitres représente 94% de la taille de la collection.

Chaque chapitre aborde un aspect différent de l'expression des documents par la grammaire TEI. Les premiers chapitres discutent des éléments communs à tous les types de documents. Les chapitres suivants discutent des éléments *facultatifs* de la grammaire, qui sont inclus selon le type de document ou les fonctionnalités désirées: par exemple, un chapitre porte sur l'expression des références internes et externes, un autre sur les éléments servant à marquer la poésie, etc.

Les chapitres ont tous sensiblement la même organisation: d'abord une discussion générale sur les particularités des informations à exprimer, ensuite la description des marques TEI permettant leur expression, illustrée par de nombreux exemples, et enfin, la définition formelle des marques et attributs TEI à travers le formalisme SGML.

5.1.2 Le formalisme TEI

Il n'existe pas de correspondance parfaite entre le formalisme d'expression originel des documents, TEI, et le langage de représentation $\mathcal{L}$. En effet, certaines informations de $\mathcal{L}$ n'apparaissent pas dans la collection ou ne sont pas exprimables dans TEI, et de même, certaines informations de TEI ne sont pas exprimables dans $\mathcal{L}$. Nous discutons maintenant de ces différences.[3]

2. L'utilisation de ces documents et les extraits reproduits dans ce chapitre, est rendue possible grâce à l'autorisation de l'ACH (Association for Computers and the Humanities), l'ACL (Association for Computational Linguistics), et l'ALLC (Association for Literary and Linguistic Computing).

3. Les éléments qui peuvent effectivement être traduits d'un langage à l'autre sont présentés à la section 5.2.2.

a) Éléments manquants dans TEI

Les informations suivantes de $\mathcal{L}$ n'ont pas d'équivalent dans notre collection:[4]

- l'incertitude et les alternatives;
- certains types, notamment les marques linguistiques *marque-ling* (les thèmes et sujets ne sont pas identifiés) et de *méta-discours*, *lexical* (les catégories lexicales ne sont pas identifées), *table* (le corpus ne contient pas de tableau), et *contenu-non-textuel* (le corpus ne contient pas d'images ou autres objets non-textuels);
- certains attributs, notamment *alphabet*, *nom-agent*, *valeur-temps*, *publié-à*, *colonnes* et *lignes*.

Deux des informations manquantes, les *intention* et les *localisation*, sont ajoutées au corpus par une reconnaissance automatique des expressions de méta-discours. Pour ce faire, le formalisme TEI est augmenté en ajoutant des sous-types à <note> (voir section 5.2.1).

b) Éléments en plus dans TEI

Les informations ayant trait à la *présentation* ne sont pas exprimables dans $\mathcal{L}$; ces informations gouvernent l'apparence du texte, comme par exemple les caractères italiques, ou la disposition physique, comme par exemple la pagination.

De même, les éléments de *documentation*, comme par exemple l'entête électronique, les annotations, ou les interventions éditoriales, ne sont pas représentées dans $\mathcal{L}$, à l'exception des attributs bibliographiques pouvant être trouvés dans l'entête électronique.

5.1.3 Restrictions sur la dérivation de thèmes

Pour des raisons pratiques, trois grandes catégories de règles ont été ignorées dans le processus de dérivation de thèmes. Les règles liées au contenu sémantique (H1) sont ignorées parce qu'elles nécessiteraient l'usage d'une base de connaissances, qui n'est pas disponible pour les *Recommandations TEI*, et qui serait trop coûteuse à réaliser pour les seuls besoins de notre expérimentation. Les règles liées à la dépendance (H2) font appel à des informations qui ne sont pas présentes dans la collection. Certes, ces informations pourraient facilement être ajoutées par une simple analyse morpho-lexicale du texte, mais encore une fois, ceci nécessite une mise en œuvre qui dépasse le cadre de cette expérimentation. Enfin, la règle de progression thématique (H4) est aussi rejetée. C'est sans doute la plus complexe à appliquer, puisqu'elle implique la reconnaissance du *sujet* et du *commentaire*.

4. L'incertitude, l'alternative, les catégories lexicales, les tableaux et leurs attributs (*colonnes* et *lignes*), sont exprimables dans TEI, mais n'apparaissent pas dans les *Recommandations TEI*

En résumé, seules les règles de dérivation suivantes ont été appliquées:

- la dérivation des thèmes (H3) par les *index* (règle 3.a) ou les marques sémantiques (règle 3.c);
- la dérivation des intentions (H5) par le titre (règle 5.a) et les expressions de méta-discours (règles 5.b et 5.c);
- la dérivation par la structure (H6) avec la règle d'héritage ascendant (6.b).

5.1.4 Architecture du prototype

Notre expérimentation a été réalisée dans le cadre de PIF, qui est un système spécialement conçu pour faciliter le développement et l'expérimentation de prototypes de systèmes de recherche d'informations. PIF permet d'organiser les différentes fonctions d'un système de recherche d'informations autour d'une même interface et de structures de données communes. La réalisation du prototype s'est effectuée par la modification des fonctions existantes de PIF et par l'ajout de nouvelles fonctions. Pour plus de détails sur PIF, voir l'annexe B.

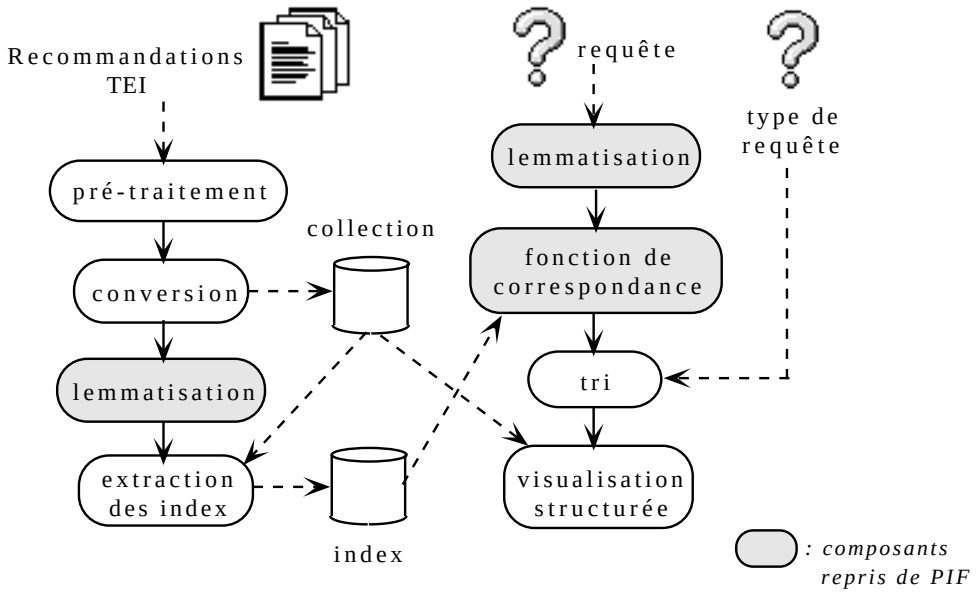


Figure 5.1. Architecture générale du prototype

La figure 5.1 présente les principales fonctions de notre prototype. Les fonctions grisées sur la figure, soient la *correspondance* et la *lemmatisation*, font partie du «*noyau*» de PIF,

et ont donc été reprises directement ou avec de légères modifications[5].

On remarque à gauche, sous les *Recommandations TEI*, les étapes du processus de représentation des documents et des index, et à droite, sous la requête, les étapes pour l'évaluation de la requête. Ces étapes sont décrites en détails dans les sections suivantes. Mentionnons tout de même les spécificités suivantes par rapport à un système traditionnel:

- une phase de pré-traitement permet d'ajouter certaines informations à la collection, notamment les expressions du méta-discours;
- en plus du fichier d'*index*, un fichier *collection* est produit, qui traduit le corpus sous un format normalisé correspondant au langage de représentation $\mathcal{L}$;
- la phase de correspondance est scindée en deux parties: on distingue l'étape de *sélection* des documents du *tri* de ces documents.

5.2 Représentation et indexation des documents

Nous avons déjà mentionné que la première fonction du langage $\mathcal{L}$ était de définir une ontologie, et qu'il pouvait être remplacé par des formalismes plus adéquats lors de l'implémentation. En l'occurrence, nous utilisons dans notre prototype le formalisme offert par PIF, qui est basé sur les *objets* (*frames*). L'application de ce formalisme à notre prototype est notée $\mathcal{L}_{PIF}$ par la suite.

Le passage de $\mathcal{L}$ à $\mathcal{L}_{PIF}$ est trivial: il s'agit de considérer un prédicat binaire $p(x, y)$ comme la propriété (*slot*) p de valeur y d'un objet x. Les prédicats unaires sont considérés comme définissant une propriété spéciale *type*; un prédicat $t(x)$ est conceptuellement équivalent à $type(x, t)$. Ainsi, l'exemple ci-dessous montre l'équivalence entre la représentation $\mathcal{L}$ (à gauche) et $\mathcal{L}_{PIF}$ (à droite):

abbréviation(*a42*).	**Passage: a42**.
texte(*a42*, «*RI*»).	TYPE: abbréviation
expan(*a42*, «*Recherche d'Informations*»).	TEXTE: RI
	EXPAN: Recherche d'Informations

$\mathcal{L}_{PIF}$ est restreint aux propriétés et aux types qui ont un équivalent dans TEI (leur liste exhaustive est donnée à la section 5.2.2).

5.2.1 Pré-traitement

Cette étape a pour but de «*préparer*» le corpus à l'indexation, en modifiant ou en ajoutant certaines informations dans les textes originaux, tout en les laissant dans le format

5. La fonction de correspondance est légèrement modifiée afin de considérer non pas des poids numériques, comme c'est le cas dans PIF, mais des couples *justification/contenu sémantique*.

TEI.

a) Modification des références

À l'origine le corpus consiste en un seul document comprenant 33 chapitres. Dans le cadre de notre expérimentation, nous considérons ces chapitres comme autant de documents indépendants. Ceci implique que les références entre chapitres soient converties en des références externes.

Dans le cas où la partie référée est un chapitre entier, les marques <ptr target=CHAP> sont converties en <xptr doc=CHAP>. Dans le cas où il s'agit de parties de chapitres, les marques <ptr target=CHAP-SEC> sont traduites en <xptr doc=CHAP from='ID(SEC)'>, où CHAP est le chapitre contenant la section SEC.[6]

b) Identification du méta-discours

Les expressions de méta-discours sont formées de trois éléments (section 4.2.5): une *localisation*, un *verbe-descriptif*, et l'*intention*. La *localisation* et le *verbe-descriptif* pour notre expérimentation sont des expressions simples du genre «*Section 3.4 describes X*» («*La section 3.4 décrit X*») ou «*X is described in section 3.4*» («*X est décrit à la section 3.4*»), données par la grammaire suivante:

`localisation` ::= ['in'] [`dét`] `dés` [`ordre`]
`dét` ::= 'this' | 'the present'
`dés` ::= 'chapter' | 'section '
`ordre` ::= '<ptr...>' | '<xptr...>'
`verbe-descriptif` ::= to describe

Comme ces informations ne sont pas prévues dans le formalisme TEI, nous les représentons en ajoutant un sous-type aux éléments <note>, noté par l'attribut type.[7] On a donc <note resp=PIF type=X>, où l'attribut resp est utilisé pour indiquer que cette information est ajouté par PIF, et où X est soit localisation ou intention. Le *verbe-descriptif* n'est pas identifié, puisqu'il s'agit toujours du verbe «*to describe*» (*décrire*).

La figure 5.2 montre un exemple de pré-traitement sur un passage TEI. L'identificateur intloc4 est ajouté afin de lier l'intention à sa localisation. En toute logique la particule «*is*» devrait accompagner «*described*» plutôt que d'être incluse dans l'*intention*; ceci provient de notre grammaire qui ne reconnaît que deux formes au verbe «*to describe*»: «*describes*» et «*described*». Cette simplification n'a pas d'impact sur l'indexation puisque les particules

6. Cette syntaxe est similaire à celle employée dans *HyTime* [NK91]. Notons que <xptr doc=CHAP> est équivalent à <xptr doc=CHAP from='ID(ROOT)'>.

7. La marque <note> sans l'attribut type garde son sens initial, à savoir une note de bas de page.

a) avant pré-traitement

The TEI header is described in chapter <ptr target=HD>.

b) après pré-traitement

<note type=intention resp=PIF target=intloc4>
The TEI header is</note>
described <note type=localisation resp=PIF id=intloc4>
in chapter <xptr doc=HD></note>.

Figure 5.2. Exemple de pré-traitement

«*is*», «*are*» et «*as*» font partie de l'*anti-dictionnaire*, et sont donc rejetées.

5.2.2 Conversion

La phase de conversion consiste à traduire les documents du formalisme TEI au format $\mathcal{L}_{PIF}$. Voici les grandes lignes de cette opération:

- *identificateurs*. Chaque objet de $\mathcal{L}_{PIF}$ est identifié de façon unique par un numéro. Les identificateurs apparaissant dans le texte TEI ne sont pas gardés dans la représentation $\mathcal{L}_{PIF}$, mais sont stockés dans une base parallèle et utilisés pour la résolution des références externes;
- *séquence*. La relation *séq* n'est pas représentée explicitement; elle est déduite de l'ordonnancement des *identificateurs*, qui sont des nombres dans $\mathcal{L}_{PIF}$;
- *composition logique*. La relation de composition logique *part* est déduite de l'inclusion des éléments dans le format TEI: un élément qui apparaît au sein d'un autre élément est considéré comme une partie de ce dernier;
- *multi-types*. Le formalisme TEI ne permet pas d'assigner plus d'un type à un même élément, puisqu'une marque doit toujours être incluse dans une autre. Ceci ne pose pas vraiment de problèmes pour les *Recommandations TEI*, sauf pour certaines *item-étiquette* qui sont aussi des *terme-technique* ou des *ident*. Ce problème n'est résolu que lors de l'indexation;
- *références*. Les *réf-interne* sont identifiées par <ptr> ou <ref> dans le texte, les *réf-externe*, par <xptr> ou <xref>. Dans le cas d'une référence interne, le lien *réf* est directement «*résolu*» – c'est-à-dire que l'identificateur TEI est converti en numéro $\mathcal{L}_{PIF}$ – à partir de l'attribut target. Les liens externes par contre ne sont pas résolus lors de la conversion; on se contente à cette étape de stocker l'identificateur référé (propriété REF) et son chapitre (propriété DOC).

– *types*. Règle générale, les marques TEI correspondent à un type $\mathcal{L}_{PIF}$, sauf lorsqu'il s'agit d'attributs bibliographiques. Les exceptions sont les *intentions* et les *localisation*, données par l'attribut type, ainsi que le type *ident-marque-déf*, qui est aussi identifié par la chaîne «*!ELEMENT*»;

– Enfin, conformément à notre modèle, les marques de documentation et de style sont ignorées.

Table 5.1. Propriétés $\mathcal{L}_{PIF}$

$\mathcal{L}_{PIF}$	$\mathcal{L}$	TEI
ABBR	$abbr(x, «y»)$	abbr=y
AUTEUR	$auteur(x, «y»)$	<author>y,<byline>y
AUTEUR-CITATION	$auteur\text{-}citation(x, «y»)$	who=y
DATE	$date(x, «y»)$	<date>y,<docdate>y
EDITE-PAR	$édité\text{-}par(x, «y»)$	<editor>y
EXPAN	$expan(x, «y»)$	expan=y
EXTERNE	$externe(x, «y»)$	déduit de entity=y'
ID	$id(x, «y»)$	<idno>y
LANGUE	$langue(x, y)$	lang=y
NIVEAU	$niveau(x, y)$	<divy>
PART	$part(x, y)$	déduit de la structure
PUBLIE-PAR	$publié\text{-}par(x, «y»)$	<publisher>y
REF	$réf(x, y)$	target=y, from=y
SELECTION	$sélection(x, «y»)$	<biblscope>y
TEXTE	$texte(x, «y»)$	texte,level1=y,level2=y
TITRE-DOC	$titre\text{-}doc(x, «y»)$	<title>y,<doctitle>y
TYPE	y(x)	<y id=x>
DOC	—	doc=y
PART-DE	$part(y, x)$	déduit de la structure
REF-DE	$réf(y, x)$	déduit de target=x

La table 5.1 résume les propriétés exprimables dans $\mathcal{L}_{PIF}$, tout en donnant leur équivalent dans $\mathcal{L}$ et dans TEI. Les trois dernières entrées dans cette table sont des propriétés

qui ont été ajoutées à $\mathcal{L}$, et qui ne sont présentes dans $\mathcal{L}_{PIF}$ que pour des raisons d'implémentation. Les liens de composition et de référence inverses, donnés par les propriétés PART-DE et REF-DE, sont utilisés pour des raisons d'efficacité. La propriété DOC, comme nous l'avons vu, sert à résoudre les liens externes.

Table 5.2. Types $\mathcal{L}_{PIF}$

$\mathcal{L}_{PIF}$	*TEI*	$\mathcal{L}_{PIF}$	*TEI*
abbréviation	<abbr>	*item-liste*	<item>
citation	<cit>,<q>,<quote>	*liste*	<list>
code	<code>	*localisation*	<note type=localisation>
desc-fig	<figdesc>	*mentionné*	<mentionned>,<socalled>
division	,<divx>	*note*	<note>
étranger	<foreign>	*paragraphe*	<p>
exemple	<eg>	*phrase*	<l>,<s>
figure	<figure>	*proposition*	<cl>
glose	<gloss>	*réf-document*	<bibl>,<biblfull>
ident	<ident>	*réf-externe*	<xref>,<xptr>
ident-attr	<att>	*réf-interne*	<ref>,<ptr>
ident-marque	<tag>	*substantif*	<w>
ident-marque-déf	déduit des <eg>	*syntagme*	<phr>
index	<index>,<keywords>	*terme-technique*	<term>
intention	<note type=intention>	*titre*	<head>
item-étiquette	<label>		

La table 5.2 énumère les types $\mathcal{L}_{PIF}$ et leur(s) équivalent(s) TEI. On remarque que la hiérarchie de types définie dans $\mathcal{L}$ est peu utilisée. Ainsi on aurait pu définir des sous-types pour *division*, en utilisant l'attribut type dans les documents TEI; mais l'usage de cet attribut dans les *Recommandations TEI* s'avérait trop sporadique et incohérent.

La figure 5.3 (a et b) donne un exemple de conversion. L'extrait consiste en une *division* formée d'un *titre* et d'un *paragraphe*. L'identificateur SG n'est pas représenté dans le format $\mathcal{L}_{PIF}$; toute référence interne à cet identificateur sera remplacée par le numéro du passage tel qu'assigné par PIF, soit #1.

a) format TEI

<div1 id=SG><head>A gentle Introduction to SGML</head>
The encoding scheme defined by these Guidelines is formulated as an application of a system known as the Standard Generalized Markup Language (SGML)... </div1>

b) format $\mathcal{L}_{PIF}$

Passage #1:	**Passage #2**:	**Passage #3**:
TYPE: division	TYPE: titre	TYPE: paragraphe
NIVEAU: 1	PART-DE: #1	PART-DE: #1
PART: #2	TEXTE: A gentle Introduction	TEXTE: The encoding scheme
PART: #3	to SGML	defined by these Guidelines...

c) après lemmatisation

Passage #1:	**Passage #2**:	**Passage #3**:
	TEXTE: gentl introduct sgml	TEXTE: encod scheme defin guidelin formul applic system known standard gener markup languag sgml

d) fichier inverse

MOT-CLE: gentl	MOT-CLE: sgml
PASSAGES: 1 H5	PASSAGES: 1 H5+H6,3 H3
CONTENU: 2.77	CONTENU: 2.08

Figure 5.3. Exemple d'indexation

5.2.3 Lemmatisation

En plus de la *lemmatisation* elle-même, cette étape concerne l'emploi d'un anti-dictionnaire[vR79, pp18–19] afin d'enlever les mots usuels. La lemmatisation est alors effectuée, selon l'algorithme de Porter [Por80], sur les propriétés «*porteuses*» d'information, c'est-à-dire qui contiennent le texte du document. Sont donc exclues de la lemmatisation les propriétés suivantes: PART, REF, PART-DE, REF-DE, TYPE, NIVEAU, LANGUE, et EXTERNE.

La figure 5.3c donne un exemple de lemmatisation. Les numéros de passages du fichier converti et du fichier lemmatisé sont les mêmes; les passages qui ne contiennent pas d'entrée lemmatisée sont laissés blancs (c'est le cas du passage #1 à la figure 5.3).

5.2.4 Extraction des index

L'extraction des index procède en deux étapes: d'abord la reconnaissance et l'indexation des passages d'indexation minimaux, à partir de leurs entrées lemmatisées, et ensuite la remontée de ces index à leur parent par *héritage ascendant*.

Pour des raisons évidentes d'efficacité, la relation *thème* est inversée; plutôt que d'associer un thème à un passage d'indexation, on associe l'ensemble des passages d'indexation à un thème. Cette approche est tout à fait similaire à la construction de *fichiers inverses*.

Ce point est illustré à la figure 5.3d, où la propriété PASSAGES est la liste de tous les passages indexés par le mot-clé. Une justification accompagne chaque passage; Par exemple, la justification du mot-clé «*gentl*» pour le passage #1 est donnée par H5. Dans le cas où plusieurs règles permettent de dériver un même thème pour un passage, les différentes justifications sont séparées par le symbole «+». Ainsi, le mot-clé «*sgml*» possède deux justifications pour le passage #3: H5 et H6. La mesure de représentativité pour un passage est retrouvée en combinant sa justification et sa valeur de contenu sémantique, donnée par la propriété CONTENU. Dans notre exemple, la représentativité du terme «*sgml*» pour le passage #1 est: <H5+H6,2.08>.

a) Indexation des passages minimaux

D'un point de vue théorique, notre approche permet de dériver les thèmes pour n'importe quel passage du texte, qu'il s'agisse d'un groupe nominal ou d'un chapitre entier. En pratique, cependant, l'indexation doit être guidée par les passages du texte susceptibles d'intéresser l'utilisateur. Ainsi, il n'est pas très utile de retourner – et donc d'indexer – un groupe nominal. Nous définissons donc les *passages d'indexation minimaux* , comme étant les passages en deçà desquels il n'est pas utile d'indexer. Dans notre expérimentation, ils sont définis comme les passages de type *paragraphe*, *division* ou *document*.

Les thèmes provenant directement du texte d'un passage d'indexation minimal ont pour justification H3. Les thèmes en deçà des passages d'indexation minimaux ne sont pas

indexés pour le passage où ils apparaissent, mais uniquement pour le passage d'indexation minimal dont ils font partie. Ils prennent pour justification H3-X, où X correspond au type du passage où ils apparaissent. Des justifications multiples peuvent survernir à ce niveau dépendamment de la «*profondeur*» du thème par rapport à son passage d'indexation minimal: par exemple, soit un thème qui est marqué comme *item-étiquette*, et qui apparaît au sein d'une *liste*, elle-même comprise dans un *paragraphe*. La justification pour ce thème serait: H3-item-étiquette + H3-liste.

Les *intention*s sont thèmes pour le passage auquel elles sont rattachées avec la justification H5. Pour les titres, ce passage peut être déduit directement de la structure. Ainsi, à la figure 5.3, les mots-clés du passage #2 sont tous thèmes du passage #1. Dans le cas des expressions de méta-discours, le passage est déduit à partir de la *localisation* à laquelle l'*intention* est liée; en effet cette expression devrait normalement contenir un pointeur sur le passage. Malheureusement, ce n'est pas toujours le cas dans les *Recommandations TEI*, aussi y ajoutons-nous la reconnaissance des expressions «*this section*» et «*this paper*», qui réfèrent respectivement à la division ou au chapitre courant.

b) Indexation des passages supérieurs

Puisqu'en pratique le critère d'héritage ascendant – un thème est index d'une section s'il est index pour toutes ses sous-sections – est trop contraignant, nous appliquons plutôt un héritage ascendant *partiel*. La notion de *seuil minimal* est définie comme suit: il s'agit de la proportion des sous-parties nécessaires pour indexer le thème. Un seuil de 0% signifie que l'héritage ascendant se fait dans tous les cas, tandis qu'un seuil de 100% signifie qu'il ne se fait en aucun cas. Un seuil de 50% signifie que l'héritage se fait si plus de la moitié des sous-parties contiennent le thème.

Nous distinguons deux types d'héritage ascendant, auxquels sont associés deux seuils: *complet* et *partiel*. Lorsque la proportion des sous-parties dépasse le seuil `complet`, les thèmes sont dérivés avec la justification H6. Si la proportion des sous-parties est inférieure ou égale à `complet`, mais supérieure à `partiel`, alors les thèmes sont dérivés avec la justification H6-partiel. Dans notre expérimentation, les seuils ont été choisis de manière *ad hoc* : le seuil `complet` est fixé à 75%, et le seuil `partiel`, à 50%.

Ce processus d'héritage ascendant est aussi appelé *remontée de termes* dans la littérature. À la différence d'approches comme IOTA, où le poids des passages supérieurs est fonction du poids des passages inférieurs, dans notre approche, la représentativité des passages inférieurs n'est pas considérée lors de la remontée des termes. Lors de l'interrogation, si la représentativité de ces passages est suffisamment élevée, alors elles apparaîtront dans la liste de réponses indépendamment de leur passage parent.

c) Calcul du contenu sémantique

Nous avons vu ci-dessus comment dériver les thèmes avec leur justification: nous voyons maintenant comment calculer la seconde composante de la représentativité d'un thème, la mesure de *contenu sémantique*. Cette mesure est inspirée de la théorie de l'information sémantique de Bar-Hillel [BH64]. Le contenu sémantique d'un thème t est donné par:

$$inf(t) = -\log m(t) = -\log\left(\frac{\text{fréquence} - \text{totale}(t)}{\text{taille} - \text{corpus}}\right) = \log\left(\frac{\text{taille} - \text{corpus}}{\text{fréquence} - \text{totale}(t)}\right)$$

où $m(t)$, la probabilité du thème t, est fonction du rapport entre le nombre d'occurences de t dans la collection, fréquence-totale(t), et le nombre d'occurences de tous les thèmes dans la collection, taille-corpus. Cette mesure tend vers 0 lorsque fréquence-totale(t) est proche de taille-corpus, c'est-à-dire que le thème n'est pas assez *discriminant*, et tend vers log(taille – corpus) lorsque fréquence-totale(t) est petit, c'est-à-dire que le thème n'est pas assez fréquent.

Puisqu'il s'agit d'une mesure de représentativité globale, elle est relative à un thème et une collection donnée. Ainsi, à la figure 5.3d, la propriété CONTENU donne les valeurs de contenu sémantique pour les mots-clés «*gentl*» et «*sgml*» de 2.77 (log(16/1)) et 2.08 (log(16/2)), respectivement.

5.3 Évaluation de requêtes

L'évaluation de la correspondance entre les requêtes et les documents se fait en deux étapes.[8] Premièrement, dans la phase de *correspondance*, on sélectionne les documents qui répondent à la requête. Ensuite, dans la phase de *tri*, la liste de documents est ordonnée d'après leur mesure de pertinence. Cet ordre peut dépendre du type d'utilisateur, de son type de requête, de son interaction précédente avec le système, etc. Nous avons choisi dans notre expérimentation de baser le tri sur les *types de requête* suivants:

- *passage*. C'est la requête classique, où l'utilisateur cherche un passage du document, c'est-à-dire un paragraphe, une section, un chapitre, etc. Par exemple, chercher «*de la documentation à propos de* SGML»;

- *référence*. Une référence ou une discussion à propos d'un autre ouvrage, pas nécessairement présente dans la collection. Par exemple, «*quels sont les autres travaux qui traitent de* SGML»;

8. Avant la correspondance les requêtes sont bien sûr *lemmatisées*, par un processus tout à fait similaire à celui des documents, sauf qu'on reconnait ici les éléments propres au langage de requête: c'est-à-dire les poids et les opérateurs booléens.

- *exemple*. Un exemple d'usage d'un mot, d'une expression, ou pour les *Recommandations TEI*, d'une marque. Par exemple, «*un exemple d'utilisation de la marque* <*term*>»;

- *définition d'un terme*. La définition d'un acronyme ou d'un terme technique. Par exemple, «*qu'est-ce qu'une marque*» ou «*que signifie* SGML»;

Ces types de requêtes sont génériques, c'est-à-dire qu'ils sont applicables à toute collection. Étant donné les informations spécifiques présentes dans les *Recommandations TEI*, nous avons cru intéressant de définir deux types additionnels pour permettre de retrouver la *description* et la *définition* d'une marque:

- *description d'une marque*. À quoi sert une marque, dans quel contexte est-elle utilisée, etc. Par exemple, «*quel type d'information est représenté par la marque* <*term*>»;

- *définition d'une marque*. Quelle est la définition SGML d'une marque. Par exemple, «*quelles autres marques peuvent être contenues dans un élément* <*term*>».

Ces deux derniers types sont donc spécifiques aux *Recommandations TEI*; ils peuvent aussi être considérés comme des spécialisations du type *définition de terme*.

5.3.1 Fonction de correspondance

La phase de correspondance est réalisée par une fonction booléenne, pouvant comporter les opérateurs habituels de *disjonction*, *conjonction* et *négation* .[9] Cette fonction est reprise de PIF et emprunte donc la même syntaxe (voir la section B.1 en annexe).

À la différence des fonctions de correspondance traditionnelles, cependant, au lieu de manipuler des poids numériques, on manipule des couples *justification/contenu sémantique*. La combinaison de ces couples se fait comme suit:

- Toutes les justifications sont conservées. La justification pour un document est donnée par une liste de justifications «*simples*», séparées par le symbole «+»;

- Les valeurs de contenu sémantique sont combinées en prenant le maximum pour une disjonction, le minimum pour une conjonction, et le complément pour une négation.

Soit par exemple un document indexé par le terme «*gentl*» avec une représentativité <H5,2.77>, et par le terme «*sgml*» avec une représentativité <H6,2.08>. Pour la requête «*gentl et sgml*», suite à la phase de correspondance, ce document aurait pour représentativité <H5+H6,2.08>.

9. Il est montré dans [Par95b] comment notre modèle d'indexation s'intègre dans le modèle booléen.

Table 5.3. Résultats pour la requête «#OR (morphological lexical)»

a) Après correspondance (extrait)

rang	#passage	représentativité
3	5691	<H6-partiel,9.28>
13	5805	<H6-partiel,9.28>
19	5103	<H6-partiel,8.02>
20	6287	<H5+H6,8.02>
22	5107	<H5,8.02>
26	6362	<H5+H6,8.02>
32	6392	<H6-partiel,8.02>
40	6460	<H6-partiel,8.02>
45	6389	<H6,8.02>
52	21549	<H6,8.02>

b) Après tri (10 meilleurs)

rang	#passage	pertinence	texte
1	5107	rouge-<H5,8.02>	Both typographically and structurally...
2	6287	rouge-<H5,8.02>	Typographic and Lexical Information...
3	6362	rouge-<H5,8.02>	Lexical View
4	6389	orange-<H6,8.02>	Retaining Both Views
5	21549	orange-<H6,8.02>	Exceptions in the WSD
6	5691	vert-<H6-partiel,9.28>	Grammatical Information
7	5805	vert-<H6-partiel,9.28>	Translation Equivalents
8	5103	vert-<H6-partiel,8.02>	Print Dictionaries
9	6392	vert-<H6-partiel,8.02>	Using Attribute Values to Capture...
10	6460	vert-<H6-partiel,8.02>	Recording Original Locations ...

La table 5.3a présente un exemple de résultat après la phase de correspondance, pour la requête «#OR (morphological lexical)» («*morphologique*» ou «*lexical*»). Seules 10 des 59 réponses sont montrées dans cette table, avec leur rang et leur représentativité.[10]

10. Les documents, en apparence désordonnés, sont en réalité ordonnés par la valeur du contenu sémantique, puis de leur ordre d'apparition dans la collection (ie. par numéro de passage).

Il existe une interface Web qui permet d'interroger le système.[11] La figure 5.4 montre un exemple d'utilisation de cette interface pour la requête «#OR (morphological lexical)». L'interface permet de choisir le type de requête – ici, un *passage* – de même que le nombre maximal de réponses souhaité.

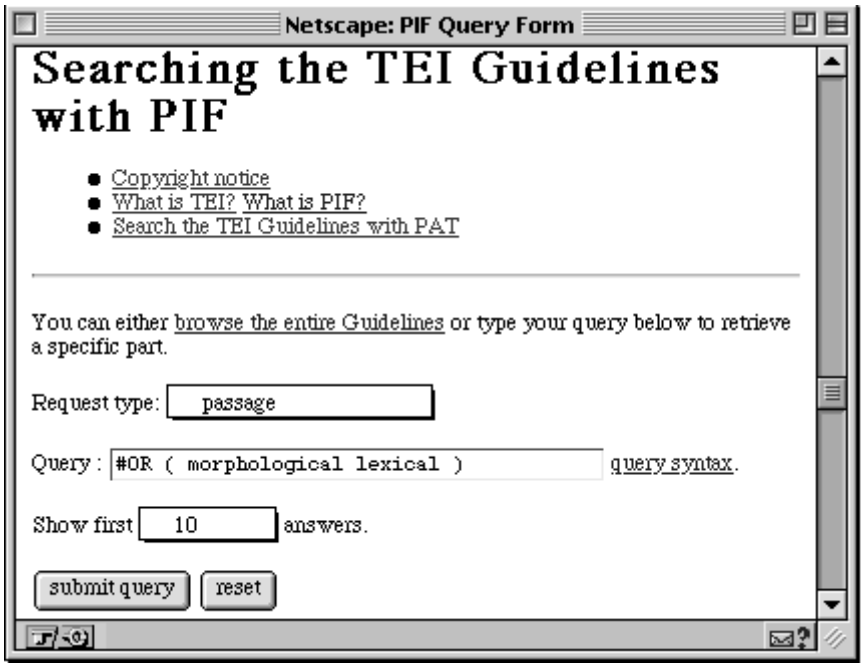


Figure 5.4. Interface pour la formulation des requêtes

5.3.2 Tri des réponses

La phase de tri reprend la liste de documents retournée par la fonction de correspondance, et les classe en ordre décroissant de représentativité. Puisque la représentativité est un couple formé de deux éléments, la *justification* et le *contenu sémantique*, le classement se fait d'abord par la *justification*, et en cas d'égalité, par le *contenu sémantique*.

Chaque type de requête est munie d'une relation d'ordre sur les justifications. Nous décrivons ci-dessous ces différents ordres.

– *passage*. La priorité est donnée aux justifications qui peuvent indiquer le thème principal d'un passage. On privilégie essentiellement les *intentions* et la *structure*. L'ordre des justifications est donc le suivant:

11. `http://www-clips.imag.fr/mrim/francois.paradis/PIF/pif.tei.html`.

H5 > H6 > H6-partiel > H3-*[12] > H3;

– *référence*. On n'est intéressé ici qu'aux justifications qui désignent un ouvrage: les *références* et – à la limite – les *intentions*. Les autres justifications sont ignorées.

H3-REF-DOCUMENT > H5;

– *exemple*. Les exemples sont aussi identifiés dans les *Recommandations TEI* par du *code*.

H3-EXEMPLE > H3-CODE > H5 > H6 > H3-*[13], H3 > H6-partiel;

– *définition d'un terme* . On recherche ici les *terme-technique*, qui, lorsqu'ils sont marqués comme tels un texte, sont souvent suivis d'une définition. Ceci est aussi vrai dans les *Recommandations TEI* pour les *étiquettes* dans les listes. Enfin, les *gloses* sont des indications directes d'une définition.

H3-TERME-TECHNIQUE, H3-ITEM-ETIQUETTE, H3-GLOSE > H5, H6 > H3* [13], H3;

– *description d'une marque*. Ces informations sont généralement indiquées par des *étiquettes* de liste, des *identificateurs* ou du *code*.

H3-ITEM-ETIQUETTE > H3-IDENT-MARQUE, H3-IDENT, H3-IDENT-ATTR, H3-CODE > H5, H6 > H3* [13], H3;

– *définition d'une marque*. Ce type de requête peut être satisfait de façon quasi parfaite par les éléments *ident-marque-déf*.

H3-IDENT-MARQUE-DEF > H3-CODE, H3-IDENT-MARQUE, H3-IDENT-ATTR > H6.

Cette phase accomplit plus que la simple réorganisation des documents; elle élimine aussi les documents qui sont jugés non-pertinents pour un type de requête donné. Par exemple, pour les trois derniers types de requête ci-dessus, les documents ayant pour justification H6-partiel sont rejetés.

La table 5.3b montre les résultats après tri, toujours pour la même requête, selon le type *passage*. Les passages #5107, 6287, et 6362, qui étaient classés respectivement au 22^{e}, 20^{e} et 26^{e} rang après la phase de correspondance (figure 5.3a), sont maintenant les trois premiers après la phase de tri. Ceci est dû à leur justification, H5, qui signifie qu'ils sont indicateurs d'intention. Les passages possédant la même justification sont ordonnés d'après leur contenu sémantique: ainsi à la table 5.3b le passage #5805 apparaît avant le passage #5103 parce que son contenu sémantique est plus élevé. Enfin, les passages ayant la même

12. La justification H3-* regroupe toutes les justifications obtenues par la règle de *marques sémantiques* (règle 3.c).

13. Ici H3-* regroupe les justifications issues de la règle de *marques sémantiques*, et non présentes explicitement dans le classement. Quant à la virgule, elle indique que les éléments ont le même classement.

justification et le même contenu sémantique appartiennent à la même *classe de pertinence*, et sont présentés à l'utilisateur d'après leur ordre d'apparition dans la collection.

La phase de tri peut en théorie identifier un grand nombre de classes de pertinence dont l'utilisateur ne saisira pas toutes les nuances, d'autant plus qu'elles ne correspondent pas nécessairement à son propre jugement de pertinence. Nous distinguons donc les classes de pertinence calculées par le système, qui sont données par les couples *justification/contenu sémantique*, de celles retournées à l'utilisateur, qui regroupent les classes de pertinence-système en cinq *niveaux* ou *couleurs*: «*rouge*» (pertinence maximale), «*orange*», «*vert*», «*blanc*» et «*gris*» (pertinence minimale). Ce regroupement a été déterminé de façon ad hoc, uniquement sur la base des justifications, et varie donc selon le type de requête. Pour le type *passage*, la correspondance est la suivante: rouge: H5, orange: H6, vert: H6-partiel, blanc: H3-$*$, et gris: H3. D'autres types, comme *référence*, n'utilisent pas toutes les couleurs.

La colonne *pertinence* dans la table 5.3b combine ces deux mesures de pertinence-utilisateur et de pertinence-système. Quant à la colonne *texte*, il s'agit d'un court descriptif du passage. qui l'accompagne lors de la visualisation des réponses.

5.3.3 Visualisation des réponses

La phase de *visualisation*, comme son nom l'indique, consiste à afficher la liste triée des réponses, et de permettre à l'utilisateur de consulter le texte intégral des passages. On ne saurait trop insister sur l'importance de la visualisation et de son impact sur le jugement de pertinence de l'utilisateur: certaines approches la placent même au cœur du processus de recherche [Sma94]. Ce critère est pris en compte dans notre prototype d'abord par une visualisation plus intuitive des degrés de pertinence, et ensuite par un affichage qui reflète la structure logique des documents.

Une question se pose quant à l'ordre de présentation des réponses: doit-il suivre strictement le degré de pertinence, ou doit-il aussi refléter la structure des documents? Nous choisissons un compromis entre ces deux solutions, où la structure logique est toujours explicitée, quel que soit le degré de pertinence des passages, mais où la pertinence l'emporte sur la structure séquentielle ou l'ordre d'apparition des passages dans les documents. Cette mise en évidence de la structure logique est réalisée en présentant d'abord les passages qui en contiennent d'autres, et en décalant ces derniers par rapport à leur parent.

Un exemple d'interface de visualisation est présenté à la figure 5.5, toujours pour la requête de *passage* «*#OR (morphological lexical)*». Il s'agit des mêmes réponses après tri que celles présentées à la figure 5.3b, mais réorganisées de façon à expliciter les liens de composition logique entre les passages. Ainsi le passage #5107 («*Both typographically...*») apparaît après le passage #5103 («*Print Dictionaries*»), même s'il est plus pertinent, puisqu'il en est une sous-partie. Par contre, la séquentialité des passages n'entre pas en compte. Le passage #5691 («*Grammatical Information*») est affiché après le passage #6287 («*Typographic and Lexical Information in Dictionary Data*»), malgré le fait qu'il apparaisse

en premier dans le document.

L'icône sert à informer l'utilisateur du degré de pertinence d'une réponse, en prenant l'une des différentes couleurs retournées par le tri: le rouge, l'orange, le vert, le blanc ou le gris. De plus, en cliquant sur cet icône, l'utilisateur obtient le couple de pertinence pour cette réponse: c'est-à-dire la *justification* ou la règle qui a été utilisée pour le dériver, et son *contenu sémantique*. L'utilisation de la couleur permet une visualisation instantanée de la pertinence indépendamment du type de requête. Ainsi, le rouge sera associé à différentes classes de pertinence-système, selon le type de requête, mais désignera toujours l'item le plus pertinent. La couleur permet également de repérer rapidement les items les plus importants, particulièrement si ils se trouvent plus bas dans structure logique.

Le texte intégral pour une réponse est disponible par l'icône . Un exemple de passage, la 3^{e} réponse de la liste, est présenté à la figure 5.5b. Autant que possible, des styles ou des icônes sont utilisés pour démarquer les différents types d'information: l'*italique* est utilisé pour les *marques sémantiques*, l'icône pour les *notes*, etc. La navigation dans la collection est rendue possible par les icônes de remontée ( ) ou de redescente ( ), de même que par les liens de référence. Ces liens de référence sont indiqués lors de la visualisation par le titre de la section ou du chapitre référencé. Enfin, la représentation $\mathcal{L}_{PIF}$ d'un passage peut être consultée par l'icône .

5.4 Analyse des résultats

Nous discutons dans cette section des résultats de l'indexation et de la recherche sur les *Recommandations TEI*, et comparons notre stratégie par rapport à un autre système ayant aussi indexé cette collection.

5.4.1 Performances du système

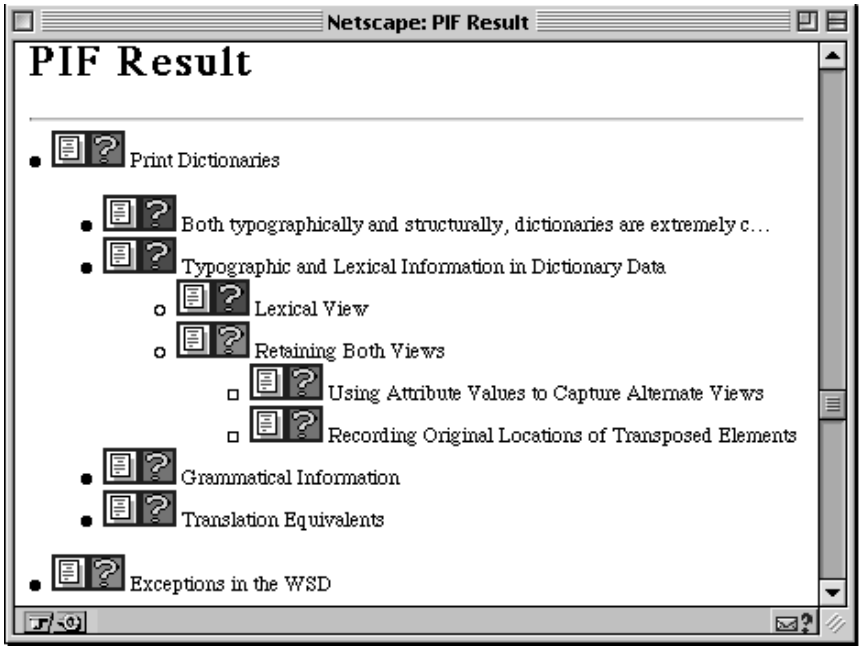
La table 5.4 présente les résultats d'indexation sur la collection des *Recommandations TEI*. La colonne *taille* donne la taille des fichiers produits, la colonne *ajout*, la différence par rapport aux fichiers source originels, et enfin, la dernière colonne donne le temps d'exécution total. [14]

Les fichiers originels font 2.23 Méga-octets. Lors du pré-traitement, 694 références externes et 248 expressions de méta-discours sont identifiées. Les *intentions* d'une centaine de sections distinctes sont ainsi identifiées [15], ce qui représente à peu près 15% de toutes les sections dans la collection. La phase de conversion découpe le texte en 21987 passages, soit une fragmentation moyenne de 666 passages par document. La collection convertie fait

14. L'expérience a été réalisée sur SUN UltraSparc 140.

15. Rappelons que les intentions ne peuvent pas s'appliquer à des passages quelconques, mais uniquement aux sections, chapitres ou documents.

a) Liste des réponses



b) Visualisation d'une réponse

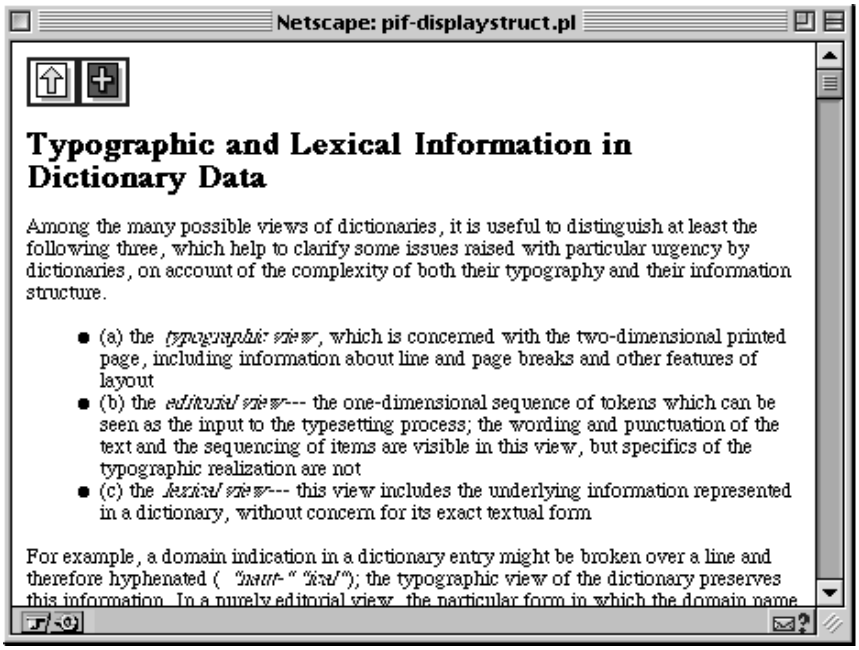


Figure 5.5. Interface pour la visualisation des réponses

Table 5.4. Résultats de l'indexation

fichier	*taille*	*ajout*	*temps CPU*
initial	2.23 Mo	–	–
pré-traitement	2.26 Mo	1 %	24 sec
conversion	2.63 Mo	15%	11 min 49 sec
lemmatisation	1.20 Mo	-86%	37 sec
indexation	1.94 Mo	-15%	2 min 37 sec
indexation-min	1.80 Mo	-24%	2 min 33 sec
indexation-max	2.83 Mo	21%	3 min 55 sec

15% de plus qu'originellement, ce qui s'explique principalement par les liens doubles et autres informations ajoutées pour des raisons d'efficacité du traitement. Après filtrage par l'anti-dictionnaire, la lemmatisation retient quelques 180,000 mots (sur un peu moins de 315,000). De ce nombre, seuls 11,689 mots-clés distincts sont conservés lors de l'indexation.

Les deux dernières lignes de la table 5.4 montrent des variantes à l'indexation. Pour «*indexation-min*», les seuils d'indexation sont fixés à 100%, ce qui signifie qu'il n'y a aucun héritage ascendant. Pour «*indexation-max*», les seuils sont fixés à 0%, ce qui signifie que tous les termes sont remontés.

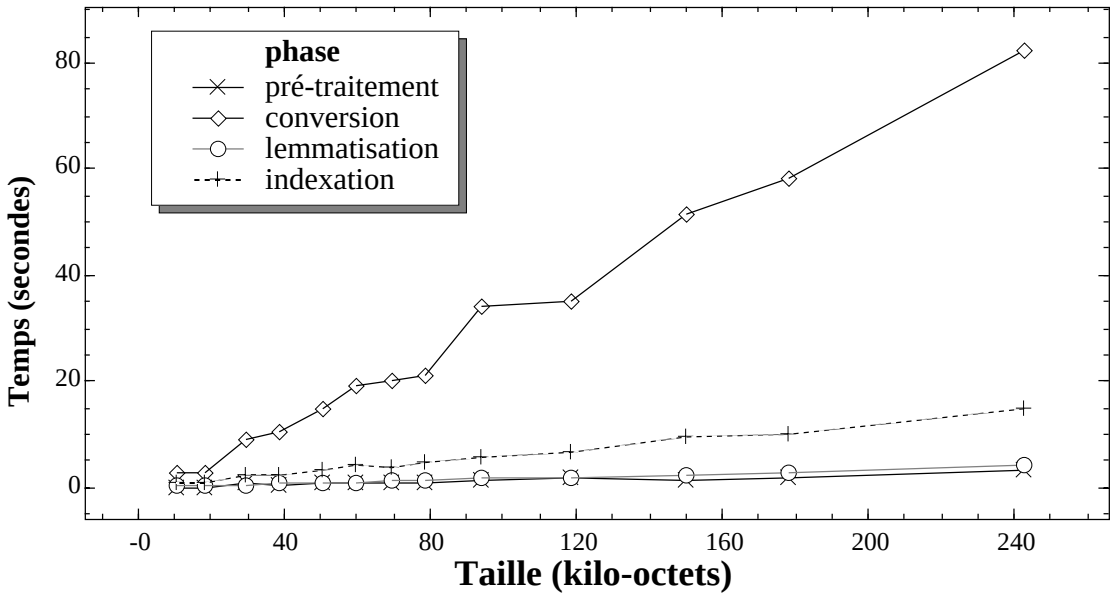
Pour ce qui est des temps d'exécution, ils sont acceptables pour une collection de cette taille. La phase la plus coûteuse en temps est la *conversion*; ceci s'explique par le fait qu'elle comporte une première étape où le texte est soumis à un analyseur SGML, et qu'ensuite il faut résoudre les références, etc.

Afin d'étudier le comportement des étapes d'indexation sur différents documents, nous présentons à la figure 5.6a les temps d'exécution en fonction de la taille des documents. Les temps d'exécution augmentent de façon linéaire par rapport à la taille: les coefficients de régression linéaire variant de 0.95 à 0.99.

La même analyse est effectuée pour la phase d'interrogation. La figure 5.6b présente les temps d'exécution moyens pour les étapes d'interrogation en fonction du nombre de documents retournés. Comme pour l'indexation, les temps d'exécution pour l'interrogation varient de façon linéaire, les coefficients de régression allant de 0.82 à 0.99.

Les temps de réponse sont plus qu'acceptables (en deçà d'une seconde) pour les requêtes comportant moins de 200 documents. Au-delà de ce nombre, l'étape de visualisation commence à ralentir le processus de façon significative, pour atteindre plus de 5 secondes pour

a) Représentation et indexation



b) Interrogation

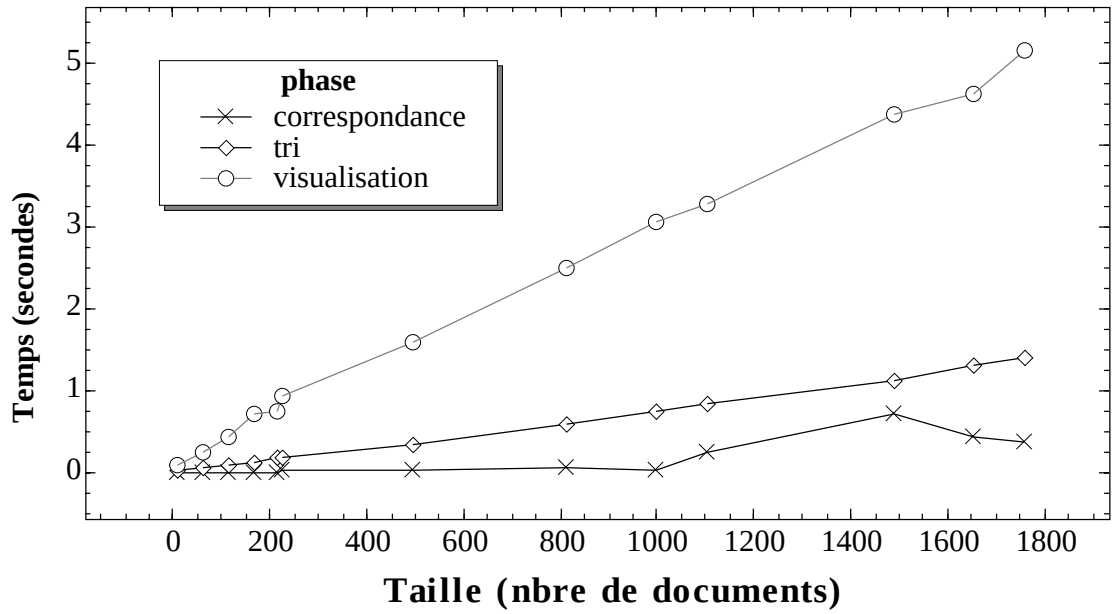


Figure 5.6. Temps d'exécution

1800 documents.[16] Il peut sembler surprenant à prime abord que la visualisation soit plus coûteuse que la correspondance ou le tri; il faut cependant considérer que cette phase requiert la lecture de la structure du document et du texte.

5.4.2 Résultats sur une requête type

A priori, notre approche devrait favoriser la précision, puisque, de façon générale, nous n'ajoutons pas de nouvelles informations à la collection, mais caractérisons l'information existante de façon plus précise. Cette hypothèse a été vérifiée en soumettant diverses requêtes au système; nous analysons ici les résultats pour l'une de ces requêtes.

La requête considérée est donnée par «*term*». Cette requête simple a été choisie à cause de l'ambiguïté de *term*, qui prend différents sens selon le type de requête. Pour un *exemple*, une *description de marque*, ou une *définition de marque*, il s'agit de l'élément TEI, pour un *passage*, il s'agit du nom commun générique, et pour une *définition de terme*, il s'agit d'un terme technique, comme par exemple «*location term*» ou «*flat term entities*».

Chaque type de requête a son propre ensemble de passages pertinents, que nous avons déterminé manuellement en parcourant plus de 1000 occurrences de «*term*» ou de mots dérivés dans la collection. Nous avons ainsi identifié 20 passages pertinents, répartis comme suit: 2 pour les types *passage*, *description de marque* et *définition de marque*, 10 pour le type *exemple*, et 4 pour le type *définition de terme*.[17]

La table 5.5 présente les résultats pour la requête «*term*». Deux approches sont comparées: dans la colonne intitulée «*avec type*» figure notre proposition, où le tri des réponses dépend des types de requêtes, alors que dans la colonne «*sans type*», le type de requête n'est pas pris en compte. La table montre le rang et la précision après chaque passage pertinent.

En moyenne, une précision de 60% est atteinte en considérant les types de requêtes, alors qu'elle n'est que de 5% en ne les considérant pas. Les meilleurs résultats sont avec les types *description* et *définition* d'une marque. Ceci n'est pas surprenant puisque ces types sont spécialement adaptés à la collection, et qu'ils font références à des informations très bien identifiées dans le texte.

Si ces résultats indiquent une augmentation significative de la précision, le rappel quant à lui est peu affecté par notre approche. Le principal élément de notre modèle susceptible d'influencer le rappel est la reconnaissance des expressions de méta-discours. En effet, lorsqu'elles font référence à des passages autres que celui où elles apparaissent, les expressions de méta-discours introduisent de nouveaux index pour le passage en question. Par exemple, pour la requête «*extending the tag set*» (*ajout à l'ensemble des marques*), le chapitre «*Modifying the TEI DTD*» (*modification à la grammaire TEI*) n'est retourné que parce que les termes de la requête figuraient dans une expression décrivant le chapitre en question.

16. Ce problème est pour l'instant contrôlé en paramétrisant le nombre maximal de réponses souhaité.
17. Le type *référence* n'a pas de réponse pour cette requête.

Table 5.5. Réponses pour la requête «term»

type de requête	*# passage*	avec type		sans type	
		rang	*précision*	*rang*	*précision*
passage	2837	3	33.3%	34	2.9%
	19400	17	11.8%	164	1.2%
exemple	2328	3	33.3%	27	3.7%
	2632	4	50.0%	33	6.0%
	2867	5	60.0%	38	7.9%
	19686	20	20.0%	175	2.3%
	19780	21	23.8%	185	2.7%
	19798	22	27.3%	189	3.1%
	19974	38	28.0%	201	3.4%
	20006	25	30.8%	206	3.9%
	20012	26	33.3%	208	4.3%
	20020	27	26.3%	209	4.8%
définition	14939	2	50.0%	102	1.0%
d'un terme	19426	3	66.6%	168	1.2%
	19429	4	75.0%	169	1.8%
	19704	6	66.6%	178	2.2%
description	2839	1	100.0%	3	33.3%
d'une marque	19472	3	66.0%	171	1.2%
définition	2871	1	100.0%	39	2.6%
d'une marque	12581	2	100.0%	82	2.4%

5.4.3 Comparaison avec PAT

Dans l'évaluation qui suit nous comparons notre approche avec PAT, un système de recherche textuelle. Si cette évaluation est certes très partielle, elle peut donner une idée des performances de notre approche.

Les *Recommandations TEI* peuvent être interrogées à l'Université de Virginie, au sein de l'*Electronic Text Center*, par le biais du système de recherche textuelle PAT [G+91] [ST92]. PAT est basé sur l'indexation de chaînes de caractères *semi-infinies* – c'est-à-dire qui commencent à un endroit donné dans le texte et qui théoriquement se poursuivent jusqu'à la fin du texte, même si en pratique on n'indexe au plus qu'une vingtaine de

caractères – qui sont classées dans des arbres de recherche appelés les *PAT trees*.

La comparaison de notre système avec PAT pose plusieurs problèmes, essentiellement dus à l'absence de la notion de structure dans PAT.[18] En effet, contrairement à notre système, PAT retourne des passages qui ne correspondent pas à la structure logique du document. Pour comparer les deux systèmes, il faut donc déterminer de façon manuelle quel paragraphe ou section correspond le mieux à un passage PAT.

Un autre problème qui découle en fait du premier est que PAT peut retourner plusieurs fois le même paragraphe ou la même section, si l'expression y apparaît plusieurs fois. Lors du calcul de précision/rappel, il faut donc considérer ces passages comme autant de documents pertinents.

Nous considérons de nouveau la requête «*term*», pouvant être exprimée selon cinq types de requête différents. La figure 5.7 présente les courbes de précision/rappel pour cette requête. Chaque courbe représente la moyenne sur les cinq types de requête considérés. Les quatre courbes correspondent aux stratégies de recherche suivantes:

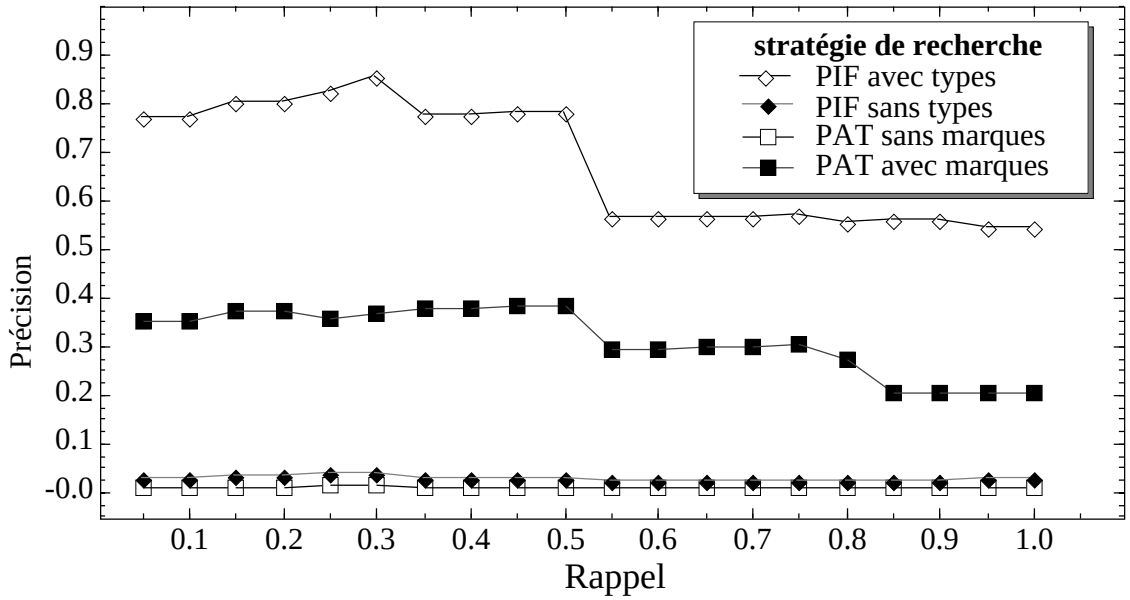


Figure 5.7. Résultats pour la requête «term»

- PIF avec types. Il s'agit de notre approche de base, qui tient compte du type de requête et du classement des justifications selon ce type;
- PIF sans types. Toujour notre approche, mais sans la notion de type de requête, c'est-à-dire que les passages ne sont ordonnés que par leur ordre d'apparition;

18. Apparamment cette notion serait bien exprimable dans PAT par le biais des *régions*, mais cette possibilité ne semble pas avoir été exploitée à l'*Electronic Text Center*.

- PAT sans marques. L'approche de base de PAT;
- PAT avec marques. On tente ici de simuler les informations qui sont utilisées dans la stratégie «PIF avec types», en reformulant les requêtes avec l'opérateur «*NEAR*» (*près de*) et les marques SGML. Par exemple, au lieu de chercher «*term*» pour une requête de type *définition de marque*, on cherche «*!ELEMENT NEAR term*», où «*!ELEMENT*» est la marque utilisée pour les définitions SGML.

Les résultats montrent un net avantage de notre approche sur PAT. Ceci est surtout dû au classement selon les types de requête. En effet, on peut constater que les résultats pour la stratégie «PIF sans types» sont similaires à ceux de «PAT sans marques». Les deux courbes sont très proches de zéro – effet qui est encore accentué par l'échelle du graphe. Ceci n'est pas surprenant, car ces deux courbes ne considèrent pas le type de requête, lequel est essentiel dans ce cas pour bien interpréter le besoin de l'utilisateur, puisque la requête elle-même est exprimée de façon assez vage par un seul terme. Le nombre de documents pertinents étant très petit (10 au plus), et le nombre de documents retrouvés assez grand (plus de 200 pour notre système, et plus de 1000 pour PAT), on pouvait s'attendre à une précision aussi faible.

L'avantage de notre approche est aussi corroboré par la courbe «PAT avec marques», qui montre que l'utilisation des marques dans le but de simuler les types de requête peut améliorer les résultats. Les courbes «PAT sans marques» et «PAT avec marques» peuvent être vues comme les limites minimum et maximum de PAT pour cette requête. Elles peuvent aussi être vues comme la différence entre un utilisateur naïf et un utilisateur expert; ce dernier, connaissant parfaitement le langage et la collection, saura en effet mieux formuler sa requête.

Bien entendu, cette requête est très appropriée pour notre approche. Le système PAT est aussi désavantagé du fait qu'il n'effectue pas de classement des passages, sur lequel reposent les mesures de précision/rappel, les réponses étant simplement présentées d'après leur ordre d'apparition dans la collection.

5.5 Conclusion

Nous avons présenté dans ce chapitre une expérimentation qui intègre dans un prototype complet notre modèle d'indexation, et qui démontre qu'il peut être implémenté de façon performante et efficace. Si le prototype a été montré ici avec TEI, il n'est en rien limité à ce formalisme; d'autres expérimentations avec des documents HTML et LaTeXont été réalisées. Cependant, chaque collection a ses types de requête particuliers.

En ce qui concerne les restrictions apportées au modèle, la règle qui pourrait être ajoutée le plus facilement est sans doute la *dépendance* (H2): il s'agirait d'ajouter au *prétraitement* une phase d'identification du type lexical. Les règles d'*équivalence* et d'*inférence* (H1) pourraient aussi être approximées par l'emploi d'un thésaurus tel que Wordnet.

Les résultats de notre approche ont été comparés au système PAT; il apparaît que le classement des justifications selon le type de requêtes semble une voie prometteuse pour améliorer la qualité des réponses. La comparaison est partielle, puisqu'elle n'a porté que sur cinq requêtes – ou plus précisément, sur une même requête, mais de cinq types différents. Avant de pouvoir réaliser des expérimentations plus complètes, il est impératif de chercher à automatiser la comparaison entre les deux systèmes: rappelons que l'analyse de la requête présentée ici a nécessité la vérification manuelle de plus de 1000 passages! Cette comparaison avec PAT fait aussi apparaître un avantage de notre système: son accessibilité à tout type d'utilisateur. Ainsi, si l'on peut atteindre des résultats similaires avec PAT par reformulation de la requête, ceci nécessite une bonne connaissance du langage de requête et de la structure interne de la collection.

Chapitre 6

Conclusion

Chercher n'est pas une chose
et trouver une autre,
mais le gain de la recherche,
c'est la recherche même.

Saint Grégoire de nysse
(*Homélies sur l'Ecclésiaste*)

6.1 Conclusion et apports

Notre travail a pour objectif la définition d'un modèle d'indexation pour les documents textuels structurés. L'importance d'un modèle formel d'indexation en recherche d'informations n'est plus à démontrer: en effet l'étape de correspondance entre les requêtes et les documents repose principalement sur la justesse et la richesse d'expression du modèle d'indexation. De plus, par rapport à un modèle *ad hoc*, un modèle formel présente l'avantage d'être indépendant du format des documents et du domaine d'application. Cependant, peu de modèles d'indexation ont exploité la richesse des documents actuels. Notre approche se veut un pas dans cette direction, en faisant de la *structure* des documents une notion centrale, et en modélisant également les *méta-informations* qui caractérisent le contenu. Ces informations sont désormais accessibles aux systèmes de recherche d'informations, alors que le monde de l'édition électronique adopte des standards comme SGML pour la représentation des documents.

La richesse des informations *explicitées* dans les documents autorise un transfert de certains problèmes de représentation réputés difficiles hors du noyau d'indexation. C'est la philosophie adoptée dans notre modèle: les problèmes classiques de traitement automatique de langue naturelle, comme la polysémie ou le paraphrasage, sont supposés résolus lors d'un processus de désambiguïsation qui précède l'indexation. Les cas ne pouvant être résolus à ce stade sont explicitement indiqués dans la représentation.

Par rapport aux modèles existants, l'originalité de notre approche se situe essentiellement à trois niveaux: quant à la *définition*, à la *représentation* et à la *dérivation* des index.

- *définition.* La notion de *thème* dans notre modèle intègre plusieurs interprétations tirées de théories linguistiques ou de discours. Ces interprétations se reflètent dans les stratégies de dérivation de thèmes, et se veulent une meilleure approximation du processus cognitif d'extraction de thème, ce processus étant trop complexe pour être réduit à une seule interprétation, comme c'est le cas avec les techniques statistiques;

- *représentation.* Un langage de représentation des documents textuels structurés a été défini, qui permet l'indépendance des règles de dérivation par rapport au formalisme d'expression des documents initiaux. Ceci s'avère crucial à mesure que les systèmes de recherche d'informations traitent des documents électroniques hétérogènes et plus riches en information. Dans notre modèle, la représentation des index se fait au même niveau que la représentation des autres informations du document servant à dériver ces index. Ceci permet notamment de ne pas cacher des informations dans les règles de dérivation, comme c'est souvent le cas dans les modèles existants;

- *dérivation.* Nous définissons des règles qui permettent la dérivation d'index à partir d'autres index ou d'informations dans le document, comme la structure, le type, les expressions de méta-discours, etc. L'importance relative de ces index n'est pas mesurée par un poids numérique, mais par une *justification* qui découle directement de la règle employée pour dériver l'index. Cette mesure de représentativité se veut plus flexible et va permettre lors de l'interrogation un classement de pertinence dépendant du type de requête;

Un autre aspect important de notre travail a trait à son évaluation: nous avons en effet validé notre approche *a priori*, en comparant l'indexation manuelle réalisée par des utilisateurs à nos propres règles de dérivation. Cette expérimentation a confirmé la pertinence de nos règles, mais a aussi démontré la grande diversité des réponses, ce qui semble suggérer qu'il ne s'agit pas de savoir quelles sont les règles de dérivation, mais qu'il faut encore savoir quand les appliquer, ou pour quel type d'utilisateur.

La validation *a posteriori*, elle, a consisté à appliquer notre approche à un corpus technique. Nous avons vu comment le modèle d'indexation s'intégrait dans un prototype complet, et comment la mesure de représentativité de thèmes pouvait permettre d'ordonner les réponses lors de l'interrogation selon le type de requête. Des résultats de comparaison avec le système PAT montrent un net avantage pour notre approche.

6.2 Perspectives

Les techniques de traitement de langue naturelle ont souvent été jugées dans le passé trop coûteuses pour être appliquées directement en recherche d'informations. En conséquence, la recherche d'informations s'est dotée de ses propres techniques, mieux adaptées à ses besoins [Sme92]: qu'on pense par exemple aux algorithmes de lemmatisation, qui sont une solution pragmatique au problème de la reconnaissance de la racine d'un mot. Les travaux de linguistique s'orientant maintenant vers de plus grands volumes de données (*large scale corpora*), ces deux disciplines pourraient dans le futur faire l'objet de développements communs. Notre modèle ne pourra que bénéficier de tels avancements, en permettant éventuellement une meilleure reconnaissance des intentions du discours, de la progression thématique, etc.

De façon plus concrète, en ce qui a trait aux extensions possibles de notre modèle, le plus important ajout concerne une meilleure intégration de la certitude et du poids dans la dérivation des thèmes. Nous avons effectivement vu le bien-fondé de ces deux mesures dans la définition du langage, mais ne les avons pas ou peu utilisées dans les règles de dérivation. Un autre ajout intéressant concerne la *progression thématique*, qui pourrait être approximée, comme pour les *intentions*, par le biais d'expressions ou de mots-liens entre les phrases ou les paragraphes. Des travaux tels que [HH76] [MH91] semblent en effet suggérer qu'il soit possible de retrouver en partie cette information par des liens lexicaux.

Notre expérimentation a soulevé bien des problèmes pratiques, ne serait-ce que pour le choix d'une collection qui regroupe les informations utilisées par notre modèle. Plus sérieux encore, elle soulève un problème théorique: celui de l'évaluation de pertinence. Les mesures de précision/rappel, que nous avons utilisées ici à défaut d'autre chose, sont inappropriées pour l'évaluation de larges collections où les documents sont de plus structurés. Une voie de recherche future consiste à concevoir une mesure de pertinence qui évalue des *passages* et non plus des *documents*, ce qui implique entre autres la reconnaissance de l'*inclusion* entre passages, de la différence de classement selon que l'on s'intéresse à l'évaluation pure ou à la visualisation, etc.

L'utilité de notre approche se situe non seulement au niveau de la recherche d'informations textuelle, mais également par rapport aux développements récents de la recherche d'informations multimédia. Dans un contexte multimédia, où l'on doit tenir compte de la richesse des documents électroniques actuels et à venir, et où les règles de dérivation doivent pouvoir utiliser au mieux les spécificités de chaque média, la notion de modèle formel d'indexation s'avère plus que nécessaire. En outre, il peut être intéressant de voir si la typologie de l'information définie ici peut être appliquée à d'autres médias. Par exemple, dans un modèle d'image tel que défini dans [Mec95], l'information est partitionnée en cinq «*vues*»: la vue *physique* correspond au contenu signifiant dans notre modèle, la vue *symbolique* au contenu signifié, la vue *perceptive* au méta-contenu signifiant, et les vues *structurelle* et *spatiale* à la structure de contenu signifiant. La structure de contenu signifié, si elle existait, correspondrait à la disposition topographique *réelle* des objets dans l'espace, par

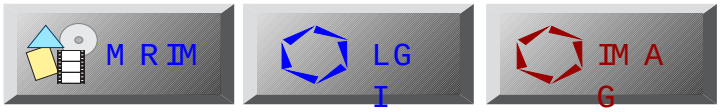
opposition à leur disposition dans la projection bidimensionnelle de l'image, qui est donnée par les vues structurelle et spatiale.

Enfin, nos travaux s'inscrivent dans un effort plus général de modélisation *contextuelle* des systèmes de recherche d'informations, qui intègre des critères jusqu'à maintenant considérés *externes* au système, ce qui permet notamment une meilleure prise en compte de l'utilisateur et de ses besoins par le jugement de pertinence [Den96]. Nous avons vu deux de ces critères dans notre expérimentation: d'une part, lors de la validation *a priori* du modèle, le type d'utilisateur, et d'autre part, lors de la validation *a posteriori*, son type de requête. Dans une telle optique, l'indexation ne doit plus être considérée comme une association statique entre des termes d'indexation et des documents, mais plutôt, ainsi que nous l'avons proposé, comme une mise en évidence de diverses propriétés des documents, dont l'utilisation dépend du contexte de recherche d'informations. Ainsi les différentes relations d'ordre de la mesure de représentativité que nous avons proposées, sont les prémisses d'un couplage qui reste à définir entre notre modèle d'indexation et un modèle de pertinence contextuel.

Annexe A

Formulaires pour la validation-utilisateur

A.1 Formulaire français



Expérience sur l'identification des thèmes dans les textes descriptifs

Utilisez plutôt cette page avec **Mosaic**, ou Netscape pour **Mac**!!!

Introduction

Cette expérience évalue votre capacité à sélectionner les éléments d'information les plus importants dans des textes descriptifs. 15 à 30 minutes devraient être nécessaires pour compléter le questionnaire. Merci!

Identification Personnelle

Avant de commencer, veuillez répondre aux questions suivantes. Cette information restera strictement confidentielle et sera utilisée uniquement à des fins d'analyse statistique.

Nom: ☐

E-mail: ☐

Comment évaluez-vous votre maîtrise du français écrit?

- ◯ limitée
- ◯ bonne
- ◯ excellente

Avez-vous déjà utilisé un système de recherche documentaire, dans une bibliothèque ou sur le Web?

- ◯ jamais
- ◯ quelques fois
- ◯ souvent

Êtes-vous familier avec la *Recherche d'Informations*?

- ◯ pas du tout
- ◯ moyennement, par des lectures ou des cours
- ◯ travaille ou étudie dans le domaine

La questionnaire comporte deux parties. Lisez bien les instructions pour chaque partie, et n'hésitez pas à me contacter si vous avez des questions.

Veuillez noter l'heure; il vous sera demandé à la fin d'estimer le temps qu'il vous a fallu pour répondre au questionnaire.

Partie I

Instructions

Il vous est demandé de choisir parmi deux éléments lequel représente le mieux le thème ou le sous-thème d'une expression donnée. Par exemple, si "rouge" et "voiture" vous étaient proposés pour l'expression "une voiture rouge", il faudrait préférer "voiture" à "rouge", puisque l'expression est à propos d'une "voiture". Les deux éléments proposés ne sont pas nécessairement présents dans l'expression, mais peuvent y être reliés; par exemple, on pourrait aussi vous proposer "couleur" et "automobile" pour "une voiture rouge".

Pour chacune des 20 expressions, choisissez le candidat approprié, ou "également" si vous croyez que les deux candidats ont la même importance (ou aucune importance), ou "ne sais pas" si vous n'avez aucune idée. Vous pouvez aussi entrer des commentaires pour une question dans l'espace prévu à cet effet. Dans l'exemple ci-dessous, "voiture" est sélectionné.

Exemple: "une voiture rouge"

- ◯ rouge
- ◉ voiture
- ◯ également
- ◯ ne sais pas

Commentaires:

Quand les deux éléments peuvent être considérés comme thème, choisissez le plus important. Dans certains cas, le thème principal n'apparaît pas dans les choix. Dans ce cas, choisissez l'élément le plus près d'un sous-thème. Dans l'exemple ci-dessous "couleur" a été sélectionné, bien que "voiture" soit le thème principal.

Exemple: "une voiture de belle couleur"

- ◯ belle
- ◉ couleur
- ◯ également
- ◯ ne sais pas

Commentaires:

1. Une tasse de café a été laissée sur la table.
   - ◯ café
   - ◯ tasse
   - ◯ également
   - ◯ ne sais pas

   Commentaires:

2. Les voitures diesel sont généralement considérées comme étant plus dommageables pour l'environnement.
   - ◯ diesel
   - ◯ voitures
   - ◯ également
   - ◯ ne sais pas

   Commentaires:

3. La France est bien connue pour ses vins.
   - ◯ Bordeaux
   - ◯ Europe
   - ◯ également
   - ◯ ne sais pas

   Commentaires:

4. Les dauphins sont des animaux magnifiques et intelligents. Malheureusement, la pollution de l'eau les menace grandement.
   - ◯ dauphins
   - ◯ pollution de l'eau
   - ◯ également
   - ◯ ne sais pas

   Commentaires:

5. De grands esprits sont nés le onze novembre.
   - ◯ onze
   - ◯ novembre
   - ◯ également
   - ◯ ne sais pas

   Commentaires:

6. Chomolungma est le nom tibétain du Mont Everest.
   - ◯ Mont
   - ◯ montagne
   - ◯ également
   - ◯ ne sais pas

   Commentaires:

7. Le Mont Everest est le point culminant de la planète.
   - ◯ alpinisme
   - ◯ ski
   - ◯ également
   - ◯ ne sais pas

   Commentaires:

8. Mercure est relativement petite et ne possède pas d'atmosphère. Cette planète est très différente de géantes telles que Jupiter ou Saturne.
   - ◯ Mercure
   - ◯ Système Solaire
   - ◯ également
   - ◯ ne sais pas

   Commentaires:

9. Pluton est la planète la plus éloignée que nous connaissions dans le Système Solaire. C'est aussi la plus petite du Système Solaire.
   - ◯ Pluton
   - ◯ Système Solaire
   - ◯ également
   - ◯ ne sais pas

   Commentaires:

10. Les trous noirs sont formés par l'effondrement d'étoiles massives.
    - ◯ noirs
    - ◯ trous
    - ◯ également
    - ◯ ne sais pas

    Commentaires:

11. Le repas peut être suivi d'une tasse de café.
    - ◯ tasse
    - ◯ café
    - ◯ également
    - ◯ ne sais pas

    Commentaires:

12. L'Edelweiss est plutôt rare au Mont Everest.
    - ◯ Edelweiss
    - ◯ Mont Everest
    - ◯ également
    - ◯ ne sais pas

Commentaires:

13. Plus de 20 satellites gravitent autour de Jupiter.
    - ◯ satellites
    - ◯ Jupiter
    - ◯ également
    - ◯ ne sais pas

Commentaires:

14. Les baleines bleues ont été victimes d'une chasse intensive dans le passé. Une telle chasse menace toujours aujourd'hui une grande partie de la faune aquatique qui n'est pas protégée comme les baleines bleues.
    - ◯ baleines bleues
    - ◯ chasse
    - ◯ également
    - ◯ ne sais pas

Commentaires:

15. Les vins de France sont parmi les meilleurs.
    - ◯ France
    - ◯ vins
    - ◯ également
    - ◯ ne sais pas

Commentaires:

16. Aucun pays ne peut ignorer de nos jours l'économie globable qui dicte en partie ses décisions. C'est dans cet esprit que l'ALENA (Accord de Libre Echange Nord-Américain) a été signé.
    - ◯ économie globale
    - ◯ ALENA
    - ◯ également
    - ◯ ne sais pas

Commentaires:

17. Les baleines sont toujours une espèce menacée. La population de baleines bleues est cependant en augmentation.
    - ◯ baleines
    - ◯ baleines bleues
    - ◯ également
    - ◯ ne sais pas

Commentaires:

18. Paris possède beaucoup de cafés agréables avec terasse et à l'atmosphère détendue.
    - ◯ Paris
    - ◯ cafés
    - ◯ également
    - ◯ ne sais pas

Commentaires:

19. Certains pays sont montrés du doigt quant au problème de l'adoption d'une monnaie unique en Europe. L'adoption de l'écu sera vraisemblablement retardée tant que ces pays ne remplissent pas les conditions minimales.
    - ◯ certains pays
    - ◯ monnaie unique
    - ◯ également
    - ◯ ne sais pas

Commentaires:

20. CNN a rapporté que des problèmes sont survenus près de l'atoll de Muroroa hier soir. Selon CNN, des activistes de Greenpeace ont tenté de pénétrer dans la zone de sécurité établie par les militaires français.
    - ◯ Muroroa
    - ◯ activistes de Greenpeace
    - ◯ également
    - ◯ ne sais pas

Commentaires:

Partie II

Instructions

Un court texte humoristique vous est présenté. On vous demande d'entrer à la suite du texte, dans les espaces prévus à cet effet, une liste de *thèmes* relatifs à la section correspondante. Les *thèmes* sont les sujets de la section (ce dont on parle dans la section). Il peut s'agir de mots simples, groupes nominaux, ou de courtes phrases. Entrez un thème par ligne.

Il vous est aussi demandé de faire la distinction entre thèmes *principaux* et *secondaires*. Les thèmes principaux représentent les principaux sujets pour la section, alors que les thèmes secondaires sont des sous-sujets ou des sujets seulement mentionnés dans un passage mais qui ne concernent pas nécessairement toute la section.

Exemple:*
Les rennes (*Rangifer tarandus*) vivent dans les régions arctiques et sub-arctiques (principalement en Finlande, Suède, Norvège et Russie). Les éleveurs de rennes sont les Lapons qui, curieusement, n'ont aucun rôle dans l'histoire du Père Noël. Le culte des 8 rennes volants nous vient probablement de Moore au début du 19e siècle.

* Extrait traduit et édité de "Reindeer".

Thèmes pour ce paragraphe:

Thème(s) principal(aux)	Thème(s) secondaire(s)
rennes	distr. géographique des rennes Lapons Père Noël culte des rennes volants

Comme on peut le voir dans cet exemple, la sélection de thèmes est très subjective. L'exemple ne montre en fait qu'*une* des réponses possibles.

Le Père Noël: Mythes & Réalités **

Peu de légendes ont eu une influence aussi universelle que le mythe du Père Noël. Cependant, tellement d'absurdités ont été dites à son égard que nous croyons qu'il est temps de rectifier les faits pour le plus grand bénéfice des générations futures. Nous discutons dans ce court texte de deux conceptions erronées concernant le Père Noël.

1. L'emplacement du Père Noël

Il a souvent été dit que le Père Noël vit au Pôle Nord. On peut d'abord remarquer qu'il n'y a pas de rennes au Pôle Nord! (et s'il y en avait, ils ne feraient pas long feu face aux

ours polaires) Avec ce froid, il ne faut pas avoir toute sa tête pour vivre là-bas. En plus, les nuits polaires ont tendance à s'éterniser, ce qui va encore si vous aimez faire la fête, mais qui peut devenir un problème si vous travaillez de nuit ou si vos voisins sont trop bruyants.

Une alternative plus raisonable a été proposée avec Korvatunturi, en Laponie. Cette terre se trouve à l'extrémité nord de la Scandinavie, et présente l'avantage de ne délivrer ni passeports ni formulaires d'impôt; ce qui arrangerait bien le Père Noël (Imaginez un instant devoir déclarer tous ces cadeaux!) Les rennes ne poseraient plus de problèmes, mais le Père Noël devrait toujours faire avec les longues nuits.

Ayant discuté du véritable emplacement du Père Noël, nous allons maintenant examiner dans la prochaine section le "problème de distribution des cadeaux".

2. Comment fait-il?

Il est dit qu'à la veille de Noël, le vieil homme, un grand sac sur ses épaules, délivre les cadeaux aux enfants. étant donné la population mondiale de 5.7 milliard (1994), et la croissance annuelle de 1.5%, il y a environ 1 milliard d'enfants de moins de 12 ans dans le monde. On peut imaginer alors la taille du sac; il faudra évidemment bien plus de 8 rennes pour tirer le traîneau.

Encore plus problématique est la vitesse à laquelle le Père Noël devrait voyager pour livrer tous les cadeaux à temps. Supposant 3.1 enfants par foyer (basé sur la moyenne de fertilité à travers le monde), on obtient 330 millions de maisons à visiter. Si elles étaient distribuées également sur les 500 millions de km2 de la surface terrestre, la distance moyenne entre deux maisons serait de 1.39 km. Le problème de planification d'itinéraire est typiquement appelé *problème du voyageur de commerce*. Si nous ignorons ce problème et supposons que toutes les maisons sont disposées en ligne, le Père Noël devrait voyager à la vitesse de 19 million km/h pour livrer tous les cadeaux en 24 heures.

Les premiers colons américains ont bien vu ce problème, puisqu'ils ont introduit la notion de sélection: pour recevoir un cadeau du Père Noël, il faut être "sage" durant l'année. Malheureusement, même avec un taux de disqualification de 50%, il y a toujours 500 millions d'enfants, et le Père Noël doit atteindre la vitesse fulgurante de 13 millions km/h.

Plus récemment, des parents pragmatiques ont utilisé l'argument que le Père Noël a plusieurs assistants, non seulement pour fabriquer les cadeaux, mais aussi pour les distribuer. S'il pouvait diviser le travail entre 383 000 assistants, alors chacun pourrait voyager à 50km/h (une vitesse raisonable pour les rennes). Cependant, les deux questions suivantes se posent: comment faire pour s'assurer que deux Père Noël ne soient jamais vu en même temps au même endroit, et comment nourrir les 3 milliards de rennes qui seraient nécessaires au transport?

** La plupart des faits sur le Père Noël proviennent de " The Santa Claus Home Page". Les données

mondiales sont extraites de " CIA World Fact Book".

Thèmes du texte en entier: "Le Père Noël: Mythes & Réalités"

Thème(s) principal(aux) Thème(s) secondaire(s)

Thèmes de la section 1: "L'emplacement du Père Noël"

Thème(s) principal(aux) Thème(s) secondaire(s)

Thèmes de la section 2: "Comment fait-il?"

Thème(s) principal(aux) Thème(s) secondaire(s)

Finalement...

Vérifiez bien que vous avez répondu à toutes les questions (les questionnaires incomplets ne peuvent être acceptés). Si vous le pouvez, donnez un estimé du temps qu'il vous a fallu pour répondre au questionnaire: [ ] mins.

Ajoutez vos commentaires au besoin:

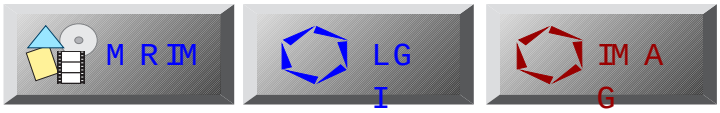
[ ]

☐ Sélectionnez cette case si vous souhaitez recevoir les résultats de l'expérience par e-mail dès qu'ils seront disponibles.

Vous pouvez maintenant [SOUMETTRE] le questionnaire.

François Paradis
23 octobre 1995

A.2 Formulaire anglais



Experiment with the Identification of Themes in Explanatory Texts

Please use this page instead if you're using Netscape for **Mac**!!!

Introduction

This experiment evaluates your ability to select the most important pieces of information in explanatory texts. 15 to 30 minutes should be necessary to complete the questionnaire. Given the anonymous and remote nature of the Web, we rely on the participants' integrity: please do not take part if you don't intend to do it seriously. Thank you!

Personal Identification

Before you proceed we would like you to answer the following questions. The information you give will remain strictly confidential and will be used only for testing.

Name: [ ]

E-mail: [ ]

How would you rate you reading skills in English?

- ◯ limited
- ◯ good
- ◯ native speaker

How often have you used a library query system or a Web search tool?

- ◯ never
- ◯ a few times
- ◯ often

How familiar are you with *Information Retrieval*?

- ◯ never heard of it
- ◯ know only from readings or courses
- ◯ working in the field

The questionnaire consists of two parts. Take some time to read the instructions for each part, and don't hesitate to mail if you have questions.

Please note the time now; you will be asked at the end to estimate how long it took you to answer the questionnaire.

Part I

Instructions

You are asked to choose which of two proposed items represent better the theme or sub-theme of a given expression. For example, if you were proposed "red" and "car" for the expression "a red car", you should prefer "car" over "red", because the expression is about a "car". The two items need not to be present in the original expression, but can be related to it in some way; for instance, you could also be proposed "color" and "automobile" for "a red car".

For each of the 20 expressions, choose the appropriate candidate, or "equally" if you think the candidates have the same importance (or no importance at all), or "don't know" if you don't have a clue. You may also input your comments for a particular question in the "comments" field. In the example below, "car" was selected.

Example: "a red car"

- ◯ red
- ◉ car
- ◯ equally
- ◯ don't know

Comments:

When both items can be considered themes, try to pick the most important one. In some cases, the main theme of the expression won't appear in the choices. In that case, select the item closest to a sub-theme. In the example below "radio" was chosen, even though "car" is the main theme.

Example: "a red car with a nice radio"

- ◯ nice
- ◉ radio
- ◯ equally
- ◯ don't know

Comments:

1. A coffee cup was left on the table.
   - ◯ coffee
   - ◯ cup
   - ◯ equally
   - ◯ don't know

   Comments:

2. Diesel cars are generally considered more harmful for the environment.
   - ◯ diesel
   - ◯ cars
   - ◯ equally
   - ◯ don't know

   Comments:

3. France is well-known for its wines.
   - ◯ Bordeaux
   - ◯ Europe
   - ◯ equally
   - ◯ don't know

   Comments:

4. Dolphins are magnificient, intelligent animals. Unfortunately, water pollution is now threatening them.
   - ◯ dolphins
   - ◯ water pollution
   - ◯ equally
   - ◯ don't know

   Comments:

5. Great minds were born on the eleventh of November.
   - ◯ eleventh
   - ◯ November
   - ◯ equally
   - ◯ don't know

   Comments:

6. Chomolungma is the Tibetan name for Mount Everest.
   - ◯ Mount
   - ◯ mountain
   - ◯ equally
   - ◯ don't know

   Comments:

7. Mount Everest is the highest place on Earth.
   - ◯ alpinism
   - ◯ downhill skiing
   - ◯ equally
   - ◯ don't know

   Comments:

8. Blind Venetians are a rare sight at Mount Everest.
   - ◯ blind
   - ◯ Venetians
   - ◯ equally
   - ◯ don't know

   Comments:

9. Pluto is the farthest planet we know of the Solar System. It is also the smallest in the Solar System.
   - ◯ Pluto
   - ◯ Solar System
   - ◯ equally
   - ◯ don't know

   Comments:

10. Black holes are formed by the collapse of massive stars.
    - ◯ black
    - ◯ holes
    - ◯ equally
    - ◯ don't know

    Comments:

11. The meal can be followed by a cup of coffee.
    - ◯ cup
    - ◯ coffee
    - ◯ equally
    - ◯ don't know

    Comments:

12. Blind Venetians are a rare sight at Mount Everest.
    - ◯ blind Venetians
    - ◯ Mount Everest
    - ◯ equally
    - ◯ don't know

    Comments:

13. Over 20 satellites revolve around Jupiter.
    - ◯ satellites
    - ◯ Jupiter
    - ◯ equally
    - ◯ don't know

    Comments:

14. Blue whales were heavily hunted in the past. Such hunting still threatens today some of the marine wildlife which is not protected like blue whales.
    - ◯ blue whales
    - ◯ hunting
    - ◯ equally
    - ◯ don't know

    Comments:

15. France's wines are among the finest.
    - ◯ France
    - ◯ wines
    - ◯ equally
    - ◯ don't know

Comments:

16. No country today can ignore the global economy which in part dictates its decisions. It was in this spirit that the NAFTA (North America Free Trade Agreement) was passed.
    - ◯ global economy
    - ◯ NAFTA
    - ◯ equally
    - ◯ don't know

Comments:

17. Whales are still a threatened species. Blue whales population, however, is on the rise.
    - ◯ whales
    - ◯ blue whales
    - ◯ equally
    - ◯ don't know

Comments:

18. Paris has a lot of pleasant, laid-back cafés with outdoor seating.
    - ◯ Paris
    - ◯ cafés
    - ◯ equally
    - ◯ don't know

Comments:

19. He didn't finish his cup of coffee. The waitress later picked it up.
    - ◯ cup
    - ◯ coffee
    - ◯ equally
    - ◯ don't know

Comments:

20. CNN reported some troubles near the Muroroa atoll last night. According to CNN, Greenpeace activists were trying to penetrate in the safety zone set up by the French military.
    - ◯ Muroroa
    - ◯ Greenpeace activists
    - ◯ equally
    - ◯ don't know

Comments:

Part II

Instructions

You will be presented a small humoristic text. Following the text are "input fields"; you are asked to enter in those fields a list of *themes* relevant for the corresponding section of the text. The *themes* are subjects for the section (what the section "talks about"); they can be single words, phrases, or even short sentences. Enter one theme per line.

You are also asked to make the distinction between *primary* and *secondary* themes. Primary themes represent the main subject(s) for that section, secondary themes are sub-subjects or things just mentioned in a passage but not necessarily recurrent in the whole section.

Example:*

Reindeers (*Rangifer tarandus*) live in arctic and subarctic regions (predominantly in Finland, Sweden, Norway and Russia). The people tending reindeer are the Lapps who, curiously enough, do not have any apparent role in Santa Claus story. The cult of flying reindeers (eight of them) was probably originated by Moore in early 19th century.

* (Edited) excerpt from Reindeer.

Themes for that paragraph:

Primary theme(s)	Secondary theme(s)
reindeers	reindeer distribution Lapps Santa Claus story cult of flying reindeers

As you can see from the example, selecting themes is highly subjective. The example was only provided to show you one *possible* answer.

Santa Claus: Truth & Myth**

Few legends have had such an overwhelming and universal influence as the myth of Santa Claus. However, so many absurdities have been said on this story that we feel it is time to set the record straight for the benefit of future generations. We shall discuss here two of the misconceptions about Santa Claus.

1. The location of Santa Claus

It has often been said that Santa lives in the North Pole. First of all, there are no reindeers in the North Pole! (and if there were, they would be no match to the polar bears) No one in his right mind would live out there, 'cause it is so damn cold. Also, northern nights tend to last forever, which is fine if you like to party, but a bit of a problem if you're working night shifts or if your neighbors are too noisy.

A much more reasonable alternative, has been proposed with Korvatunturi, in Lapland. This land lies at the northern end of Scandinavia, and has the definite advantage of issuing neither passports or tax forms; which would suit Santa well (imagine for a moment having to declare all those gifts!). The reindeers would not be a problem anymore, but Santa would still have to deal with long nights though.

Having discussed the real location of Santa's home, we will now look in the next section at the so-called "presents delivery problem".

2. How does he manage?

It is said that on Christmas Eve, the old man hauls a big sack on his back and distributes gifts to children. Given a population of about 5.7 billion (1994 figure), and an annual growth of 1.5%, there are about 1 billion children under 12 in the World. One can imagine the size of the sack then; obviously more than 8 reindeers will be necessary to pull it.

More problematic still is the speed at which Santa would have to travel to deliver the gifts on time. If we assume 3.1 children/household (based on the global fertility rate), we get 330 millions houses to visit; if they were distributed evenly throughout the 500 million km2 of Earth's surface, the average distance between the houses would be of 1.39 km. The problem of itinerary planning is typically called the *traveling salesman problem*. If we ignore that problem and suppose the houses were all laid out in a line, Santa would have to travel at 19 million km/h to dispatch all the gifts in 24 hours.

Early American settlers must have understood the catch, because they introduced the notion of selection: in order to get a gift from Santa, you need to be "good" during the year. Unfortunately, even assuming a disqualification rate as high as 50%, one still get 500 million children, and a staggering speed of 13 million km/h.

More recently, pragmatic parents have been reported using the argument that Santa has a lot of helpers, not only to manufacture the gifts, but also to distribute them. If he were to divide the job between 383 thousands helpers, then they could travel at 50km/h (a reasonable speed for reindeers). However, the following two questions immediately arise: how do you make sure no two "Santa Claus" will ever be sighted at once, and how do you feed the 3 billion reindeers that will be needed to carry them?

** Most facts about Santa were taken from The Santa Claus Home Page. World data is from the CIA World Fact Book.

Themes for the whole text: "Santa Claus: Truth & Myth"

Primary theme(s)

Secondary theme(s)

Themes for section 1: "The location of Santa Claus"

Primary theme(s)

Secondary theme(s)

Themes for section 2: "How does he manage?"

Primary theme(s)

Secondary theme(s)

Finally. . .

Make sure you have answered every question in Part I and that you have entered themes for every section in Part II (incomplete forms cannot be accepted). If you can, give an estimate of how much time you spent on the whole questionnaire: [ ] mins.

Feel free to add any comments:

☐ Check if you want results of the experiment to be e-mailed to you when they become available.

You can now [SUBMIT] the questionnaire.

François Paradis
30 August 1995

Annexe B

Le système PIF

Le système PIF est un noyau de système de recherche d'informations – c'est-à-dire qu'il comprend les composants habituels: lemmatiseur, indexeur, thésaurus, etc. – conçu de telle façon à faciliter le développement et l'expérimentation de prototypes de systèmes de recherche d'informations, en permettant différentes combinaisons, la modification, l'ajout ou le retranchement de ses composants de base. La conception de PIF répond aux besoins suivants:

- *interopérabilité des composants.* Les systèmes de recherche d'informations actuels font souvent appel à plusieurs solutions qui doivent être intégrées ensemble. Les fonctions dans PIF sont des modules indépendants, ce qui permet d'enlever rapidement un composant de la chaîne ou au contraire d'en rajouter un nouveau, par exemple pour tester les performances du système avec ou sans *thésaurus*;

- *indépendance de la collection.* La plupart des systèmes existants sont dépendants du corpus: soit directement dans leur stratégie de recherche ou d'indexation, soit indirectement parce que fortement liés au format interne de leur collection. Nous proposons dans PIF d'utiliser un format de représentation des documents intermédiaire indépendant du corpus;

- *transparence des résultats.* Il est important dans la phase d'expérimentation de pouvoir accéder et visualiser facilement tous les résultats intermédiaires, par exemple, le fichier des documents lemmatisés, le fichier inverse, etc. Ceci est assuré dans PIF par la *modularisation*, les modules ne communiquant entre eux que par fichiers ou par paramètres sur la ligne de commande. De plus, ces fichiers partageant un format commun, la visualisation s'en trouve simplifiée.

- *flexibilité.* L'utilisateur doit pouvoir modifier facilement les différentes options d'indexation et d'interrogation. Il peut s'agir de seuils d'indexation, d'anti-dictionnaire, etc. Dans PIF, chacune de ces variable est paramétrée.

B.1 Architecture de PIF

La figure B.1 montre les liens entre les principaux modules constituant le *noyau de base* de PIF, regroupés dans les phases d'indexation et d'interrogation. La «*collection*» et le «*fichier inverse*» sont des *bases* PIF, un format de base de données dans lequel sont exprimées plusieurs des données de PIF (voir la section B.3).

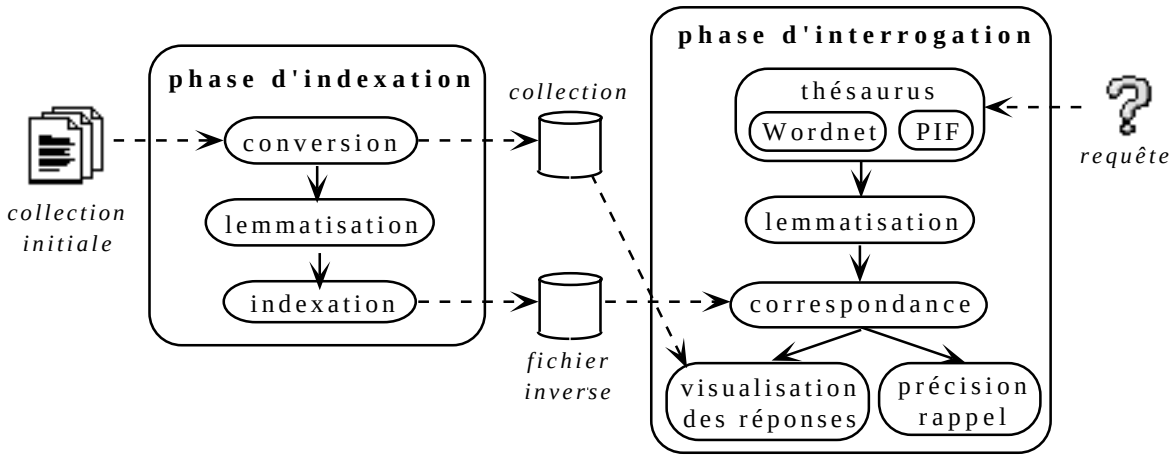


Figure B.1. Architecture générale de PIF

Nous décrivons brièvement ces modules ci-dessous. Il ne s'agit ici que de décrire dans leurs grandes lignes les fonctionnalités du système; pour plus de détails quant à leur emploi ou à leur mise en œuvre, on peut se référer à la documentation accompagnant le code source de PIF.

a) Conversion

La phase de conversion permet de traduire les documents initiaux – qui existent sous des formes très variées – en une base PIF. Pour ce faire le programme de conversion permet l'expression d'une grammaire qui dirige l'extraction des documents et de leurs champs, chaque document étant composé de champs, qui peuvent correspondre à l'auteur, au titre, etc.

La figure B.2 présente un exemple de grammaire de conversion pour la collection CACM. Il y a neuf champs possibles dans la CACM, indiqués par «*FIELD* » dans la grammaire. Le début et la fin de chaque champ sont identifiés par des chaînes de caractères dans la grammaire, respectivement par «*START* ...» et «*END* ...». En l'occurence pour la CACM

le début d'un champ est indiqué par une ligne formée d'un point («.») suivi d'une lettre – par exemple pour le champ auteur, «*.A*» – et la fin, par le début du prochain champ. Tout ce qui se trouve entre les chaînes de début et de fin dans le document initial est enregistré dans la base PIF dans un champ dont le nom est donné par «*NAME ...*», sauf les champs marqués par «*DISCARD*», qui sont tout simplement ignorés.

```
FIELD: DISCARD  // IDENTIFIER
START [.]I
END \n.

FIELD: NAME TITLE
START [.]T\n
END \n.

FIELD OPTIONAL : NAME ABSTRACT
START [.]W\n
END \n.

FIELD: NAME ISSUE
START [.]B\n
END \n.

FIELD OPTIONAL : NAME AUTHORS
START [.]A\n
END \n.

FIELD OPTIONAL : NAME KEYWORDS
START [.]K\n
END \n.

FIELD OPTIONAL : DISCARD
START [.]C\n
END \n.

FIELD: DISCARD  // TIME-ENTERED
START [.]N\n
END \n.

FIELD: DISCARD  // CROSS-REFERENCES
START [.]X\n
END \n.
```

Figure B.2. Grammaire pour la conversion de la CACM

La structure des documents de la CACM est assez stricte, en ce sens que les champs ne peuvent apparaître qu'une seule fois, et dans un ordre précis. Pour des grammaires plus complexes, on peut aussi exprimer des champs répétés plusieurs fois («*LIST:*»), des alternatives entre champs («*DISJOINT:*»), ou même des conversions de chaînes de caractères à l'intérieur d'un champ («*CONVERT ... TO ...*»).

b) Lemmatisation

La lemmatisation prend en entrée une requête (un fichier texte) ou une collection (une base PIF), et les reproduit en sortie de telle sorte que les mots vides de sens ont été enlevés et les autres termes «*normalisés*». La liste de mots vides est prise d'un *anti-dictionnaire* – par défaut, il s'agit de la liste proposée dans [vR79, pp18–19]. Deux options sont offertes pour la lemmatisation: soit le simple retrait de suffixes à partir d'une liste de suffixes types, soit la *lemmatisation* par l'algorithme de Porter [Por80].

Lorsqu'il s'agit de requêtes, les opérateurs booléens et les parenthèses sont laissés tels quels (voir la syntaxe pour les requêtes plus loin). Pour les collections, on peut aussi spécifier

quels sont les champs à lemmatiser: ainsi, pour l'exemple de la figure B.5, il ne serait pas utile de lemmatiser le champ ID.

c) Indexation

L'indexation vise la création d'un fichier inverse, qui donne la liste de documents pondérés pour chaque terme, à partir de la collection lemmatisée. Seuls sont retenus les termes dont la fréquence d'apparition dans la collection est comprise entre un seuil minimal et maximal; les termes sous le seuil minimal n'étant pas jugés assez importants pour la collection, et les termes au-dessus du seuil maximal n'étant pas jugés assez discrimants.

Tous les champs sont confondus dans le fichier inverse, c'est-à-dire que seule la référence au document est conservée. Il est toutefois possible de choisir les champs à indexer, et même de définir une pondération différente pour chaque champ. Pour l'exemple de la figure B.5, on pourrait ajouter un facteur de 2 au champ TITRE, pour traduire le fait que cet élément est plus porteur d'information que le résumé. Par défaut le poids des termes d'indexation est donné par $tf \cdot idf$ avec une pondération de 1.

d) Correspondance

La fonction de correspondance retourne une liste de documents pondérés à partir d'une requête et d'un fichier inverse. Différentes variantes de fonction de correspondance booléenne sont offertes dans PIF. L'expression des requêtes permet donc l'emploi des connecteurs logiques habituels: la conjonction, la disjonction et la négation. La syntaxe pour une requête est la suivante:

```
exp ::= terme-exp | p-exp
terme-exp ::= [ poids ] terme
poids ::= '<' nombre '>'
p-exp ::= op-exp '(' { exp }* ')'
op-exp ::= [ poids[a] ] op
op ::= '#OR' | '#AND' | '#NOT'
```

[a] Le poids sur un opérateur n'est pris en compte qu'avec la fonction de correspondance *ANDOR*.

Voici les différentes fonctions de correspondance:

- *Booléenne*. Comparaison stricte de termes. Ici les poids $tf \cdot idf$ sont ignorés, seule compte la présence ou l'absence d'un terme;
- *Booléenne floue*. Les règles suivantes s'appliquent pour la combinaison des termes: pour un #OR, on prend le maximum des poids, pour un #AND, le minimum, et

pour un #NOT, 1 moins le poids. Les poids qui accompagnent les termes dans la requête sont multipliés par le poids des documents;

- *ANDOR* [WK79]. Cette fonction est similaire à la fonction booléenne floue, sauf pour ce qui est de l'interprétation des poids sur les opérateurs dans la requête. Un nouveau connecteur, «*ANDOR*», permet des contraintes à mi-chemin entre le #AND (jugé trop strict) et le #OR (trop permissif). Ce nouvel opérateur s'exprime par des poids sur #AND et sur #OR: ainsi la conjonction stricte s'écrit «*<1.0>#AND*» ou «*<0.0>#OR*», et la disjonction, «*<0.0>#AND*» ou «*<1.0>#OR*»;

- *Bookstein* [Boo80]. Cette fonction est similaire à la fonction booléenne floue, sauf pour l'interprétation des poids dans la requête sur les termes réunis par #AND. La fonction $min(f/a, 1)$ est utilisée, où f est le poids du document indexé, et a le poids du terme dans la requête;

- *Seuils*. L'idée consiste à considérer les poids de la requête comme des *seuils*. La fonction F1 de Buell/Kraft [BK81] est implémentée. Soit f le poids du document indexé, et a le seuil dans la requête, alors le poids est donné par:

$$\begin{array}{rl} (1+a)*(f/a)/2 & \text{si} f < a \\ (1+a+a*(f-a))/2 & \text{si} f > a. \end{array}$$

e) Thésaurus

Le thésaurus permet une extension de la requête, en ajoutant aux termes des mots synonymes ou proches. Deux thésaurus sont disponibles dans PIF: Wordnet [Mil96] et un thésaurus propre à PIF.

Deux classes de *synsets* de Wordnet sont utilisées: les *synonymes*, auxquels on associe un poids de 1, et les *hypernymes* (c'est-à-dire termes plus généraux), auxqels on associe un poids de 0.8. Tous les termes du synset sont ajoutés à la requête, sans aucun effort de désambiguïsation.

Le thésaurus PIF permet d'exprimer des relations plus spécifiques à l'application, comme par exemple l'expansion des acronymes, ou la traduction anglaise de certains termes. Ce thésaurus est créé manuellement à partir d'un fichier texte qui donne les relations d'équivalence (notées par «<->») ou d'implication (notées par «->») entre les termes. Voici un exemple de tel fichier:

```
RI <-> #AND ( recherche information )
uncertainty <-> incertitude
correspondance -> <0.7>similarité
```

f) Analyse précision/rappel

Lorsqu'on dispose de la liste des documents pertinents pour une requête, on peut calculer le rappel et la précision. Il est aussi possible de calculer la précision pour un rappel donné, ce qui est utile pour la construction de courbes de précision/rappel.

g) Visualisation

La visualisation permet d'afficher le contenu d'une base PIF. Deux types de visualisation sont possibles: la visualisation *complète* ou *partielle*. La visualisation *complète* affiche tous les champs pour un numéro de document donné, ou pour l'ensemble de la base. La visualisation *partielle* affiche une représentation succinte d'une liste de documents: il peut s'agir par exemple du numéro du document suivi de son titre. Elle peut être interactive, auquel cas l'utilisateur peut se déplacer dans la liste et afficher un document particulier. La visualisation *partielle* est donc appropriée pour l'affichage des résultats d'une requête.

B.2 Interface

Deux interfaces sont proposées pour gérer l'appel des modules et le passage de données: une interface texte et une interface Web.

a) Interface texte

a) paramètres

```
corpus_stem=CACM.lem
corpus_key=CACM.keywords
freqIndMin=0.0
freqIndMax=0.50
indexer=$PIF_BIN/index
```

b) menu

```
INDEX indexer le corpus :
      $indexer $freqIndMin $freqIndMax
      $corpus_stem $corpus_key
```

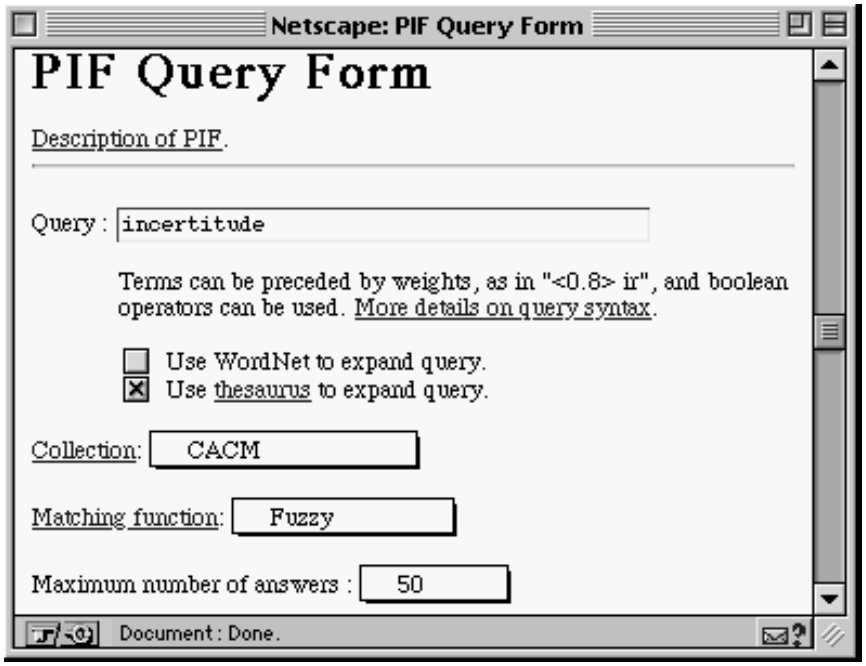
Figure B.3. Extrait des fichiers de l'interface texte pour la CACM

L'interface texte consiste en des menus qui permettent l'appel des différents modules du noyau PIF, ou des fonctions qui y ont été ajoutées. Ces menus sont configurés par le biais de fichiers de paramètres et de menus. La figure B.3 donne un exemple de configuration pour l'indexation de la collection CACM. Le fichier de paramètres contient des variables comme la fréquence d'indexation minimum (`freqIndMin`), le fichier contenant la collection lemmatisée (`corpus_stem`), etc. Le fichier de menus contient les options des menus et les

commandes qui y sont associées: ici l'option «*indexer le corpus*» est ajoutée au menu «*INDEX*», et le programme associé est donné par la variable `$indexer`.

b) Interface Web

a) Interrogation



b) Résultats

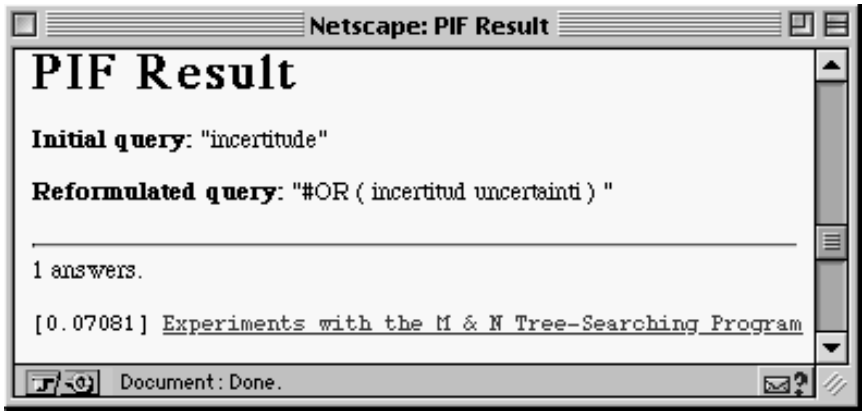


Figure B.4. Exemple d'interface Web pour PIF

Une interface Web permet la formulation des requêtes et la visualisation des réponses. La figure B.4 donne un exemple de requête pour la collection CACM. La requête est «*incer-*

titude»; l'interface permet également de choisir le thésaurus, la fonction de correspondance, ainsi que le nombre maximal de réponses. La liste de résultats – ici un seul document – montre également la requête après reformulation, c'est-à-dire une après expansion avec le thésaurus et lemmatisation. Ici, le terme anglais « *uncertainty* » a été ajouté après lemmatisation à la requête.

B.3 Les bases PIF

Les modules de PIF échangent les données dans le même format, une forme simplifiée d'objets (*frames*). Ces objets sont stockés sur disque dans des *bases* PIF, qui représentent un compromis entre la *simplicité* et l'*efficacité*: *simplicité* parce qu'elles peuvent être manipulées et visualisées facilement, et *efficacité* parce que, tout comme des bases de données, elles permettent un accès direct et efficace aux données.

Les bases PIF sont formées de trois fichiers: le fichier de format, le fichier d'index, et le fichier de données. Le fichier de données est organisé en *documents*, eux-mêmes formés de *champs*. Le fichier d'*index* permet un accès direct à ces documents en indiquant par un pointeur le début de chaque document dans le fichier de données. Enfin, le fichier de *format* décrit le nom et le type des *champs*.

Imaginons par exemple des documents pouvant contenir les champs suivants: ID (un identificateur ou numéro pour le document), TITRE (le titre du document), AUTEUR (l'auteur du document), et RESUME (un court texte résumant le document). La figure B.5 donne un exemple de représentation selon ce schéma pour deux documents identifiés par «*paradis96a*» et «*paradis94a*».

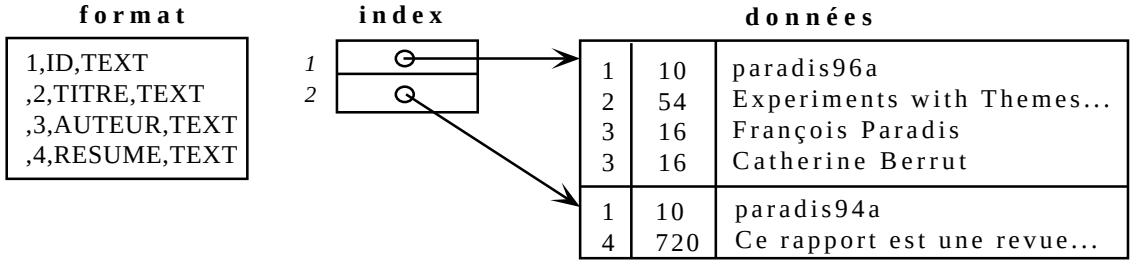


Figure B.5. Exemple de base PIF

Le fichier de format est un fichier texte où chaque ligne contient les informations relatives à un champ: un *identificateur* (nombre), un *nom* (chaîne de caractères), et un *type*. Le *type* d'un champ peut être soit TEXT ou NULL-TERMINATED-TEXT: dans les deux cas il s'agit de texte, mais dont la représentation interne dans le fichier de données diffère.

Le fichier d'index contient des pointeurs (8 octets) sur les documents dans le fichier de

données. Les documents sont implicitement numérotés, le premier pointeur correspondant au document #1, et ainsi de suite.

Le fichier de données est un fichier binaire qui stocke de façon séquentielle les champs pour chaque document. Pour chaque champ, le premier octet réfère à l'*identificateur* du champ (réf. fichier de format), suivi de son contenu, dont la représentation dépend du type de champ. Pour les champs de type TEXT, il s'agit de deux octets donnant la *taille*, et suivis du texte (une chaîne de caractères). Pour les champs de type NULL-TERMINATED-TEXT, la taille n'est pas indiquée, mais la chaîne de caractères est terminée par le caractère nul (`'\0'`). Les champs de type TEXT permettent donc un parcours plus rapide du document (puisque la taille du champ est connue à l'avance), mais leur taille est limitée à $2^{16} - 1$ octets.

Un document ne contient pas nécessairement tous les champs, et ceux-ci peuvent être répétés au sein d'un même document. Dans l'exemple de la figure B.5, le document #1 possède un TITRE, «*Experiments with Theme Extraction in Explanatory Texts*», et deux AUTEURs, «*François Paradis*» et «*Catherine Berrut*». Le document #2, lui, ne contient qu'un ID et un RESUME. Notons qu'aucun séparateur n'indique la fin d'un document et le début d'un autre, puisque cette information est implicitement donnée par le fichier d'index.

Bien que nous ayons donné ici un exemple de représentation de documents, ce format est partagé par tous les données produites par PIF: texte lemmatisé, thésaurus, fichier inverse, etc. Par exemple, pour le fichier inverse, la notion de document est remplacée par celle de mot-clé, chaque mot-clé comportant deux champs: un premier qui donne le mot-clé lui-même, et le second la liste de documents où il apparaît.

B.4 Autres applications de PIF

Le système PIF est particulièrement bien adapté au développement en parallèle par plusieurs utilisateurs. Avec l'interface texte, on peut aisément ajouter ou enlever des composants localement sans toucher au noyau principal de PIF. En plus de l'extension pour la prise en compte de la structure, telle que proposée dans cette thèse, PIF a également été utilisé au sein de l'équipe MRIM comme noyau de base d'une interface de recherche d'informations [Ant95] [MN96] et pour un algorithme d'extraction de cliques [Gau96].

Une autre application intéressante de PIF est son utilisation pour la gestion d'une bibliographie. Ainsi, les publications de l'équipe MRIM ont été indexées avec ce système et l'interrogation rendue possible par le Web, après adaptation de l'affichage des résultats pour tenir compte des spécificités des notices bibliographiques. L'interface Web a également été augmentée pour permettre l'ajout, la modification ou la suppression de publications (voir figure B.6).

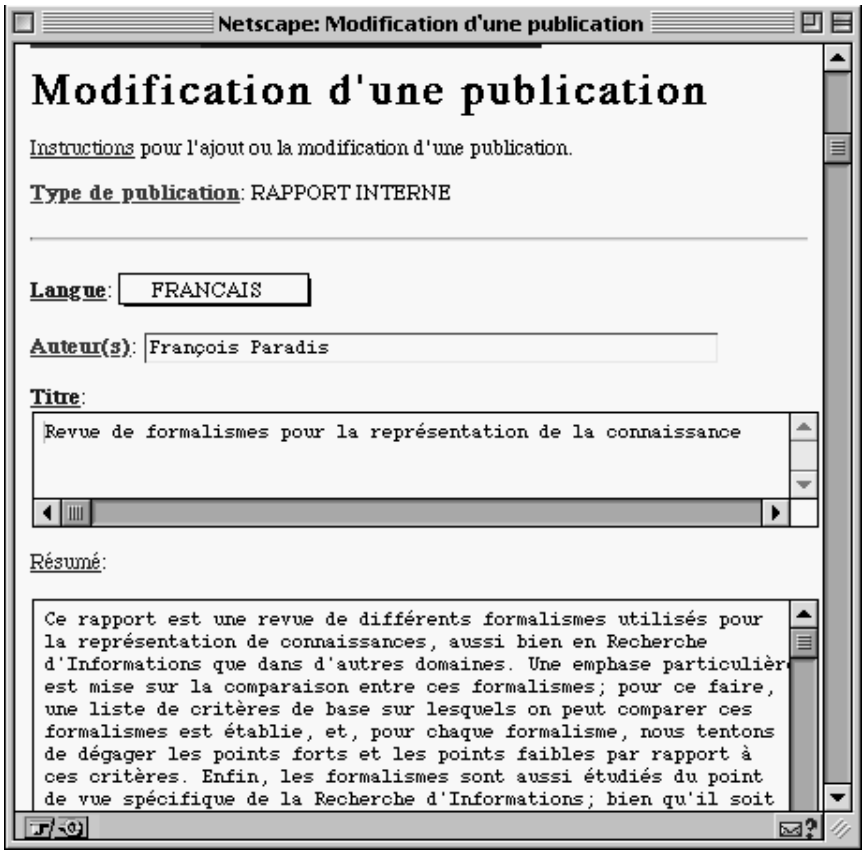


Figure B.6. Exemple de modification d'une publication

Annexe C

Les recommandations TEI

Le formalisme TEI (pour *Text Encoding Interchange*) réfère à un ensemble de recommandations pour l'encodage du texte, de ses styles et de ses différentes caractéristiques. Par rapport à un standard, ces recommandations sont plus flexibles: bien qu'un formalisme unique soit proposé, plusieurs des informations représentées sont optionnelles ou peuvent être encodées de façons différentes.

Les recommandations TEI ont été mises en œuvre par l'ACH (*Association for Computers and the Humanities*), l'ACL (*Association for Computational Linguistics*), et l'ALLC (*Association for Literary and Linguistic Computing*). Elles s'adressent donc à une très grande variété de styles de texte, incluant entre autres: le discours naturel, dans sa forme écrite ou parlée, la poésie, la prose, le théâtre, les dictionnaires, les données hypermédia, etc.

Nous ne donnons ici qu'un bref aperçu du formalisme et de ses possibilités. Une introduction plus complète peut être trouvée dans [BSM94], ou en référant au document originel des recommendations TEI, aussi connu sous le nom de «*P3*» [TEI94].

La grammaire TEI

Les recommendations TEI sont exprimées à l'aide de SGML (pour *Standard Generalized Markup Language*) [Gol91], un formalisme abstrait permettant l'expression de *schémas de marquage*, c'est-à-dire, de règles pour l'*annotation* ou l'ajout de *marques* dans un texte. Bien qu'il origine du monde de l'édition électronique, SGML peut être vu plus généralement comme un format pour le *traitement* des documents. Par opposition à des formats comme ODA, qui sont principalement destinés à être lus, corrigés, imprimés par des humains, le format SGML est lui destiné à l'échange et au traitement par des machines.

SGML est en fait un *méta-langage*, qui, par le biais d'une DTD (*Document Type Definition*), permet la définition d'un schéma de marquage spécifique. La DTD est une grammaire

qui spécifie la syntaxe et les combinaisons possibles des instructions de marquage dans les documents; elle peut aussi être vue comme la définition d'une *classe* de documents. Toutefois, elle ne spécifie pas la *sémantique* des codes de marquage, mais seulement leur *syntaxe*. C'est pourquoi SGML est dit indépendant de l'application: il ne fait que décrire le contenu, sans faire aucune supposition sur la façon dont il devrait être traité.

Les recommandations TEI sont implémentées par une DTD *modulaire*, c'est-à-dire qu'à un noyau de base commun à tous les documents (le *core tag set*), s'ajoutent des modules définissant le type de document (le *base tag set*), ainsi que d'autres composants définissant des fonctionnalités additionnelles pouvant apparaître dans plusieurs types de documents (*additional tag set*).[1] Par exemple un document de type «*prose*» pourra faire appel à un module additionnel pour la représentation de l'incertitude, s'il provient d'une source manuscrite dont la transcription était floue. Un document de type «*parlé*» fera appel au même module d'incertitude si la qualité sonore de l'enregistrement laisse à désirer.

Types d'information dans TEI

Le principe de base de l'encodage dans TEI consiste à ajouter des *caractéristiques* aux passages du texte par le biais de marques ou d'attributs SGML. La marque sert souvent à définir la nature des informations. Soit par exemple <numvalue=42>quarante-deux</num>, où num est une marque, et value un attribut. Le passage «*quarante-deux*» est ici défini comme un nombre (num); on spécifie de plus sa valeur numérique, 42.

On distingue quatre grands types d'information dans TEI: les informations de structure, de contenu, de disposition et de documentation.

– Les informations de *structure* concernent la décomposition logique du texte et les références à d'autres passages. La structure est donnée dans TEI par l'utilisation de marques SGML pour les chapitres, paragraphes, listes, etc., et par l'inclusion récursive de ces marques entre elles. Par exemple, le passage ci-dessous définit un chapitre, ou une division ( ) de type chapitre (chapter), possédant un titre (<head>), et un paragraphe de texte ( ).

  <head>Chapitre 1</head>
  La problématique de la Recherche d'Informations peut être vue comme la satisfaction d'un besoin en information d'un utilisateur

  Une référence au chapitre 1 pourrait être donnée comme suit:

  Nous avons vu <ref target=c1>au chapitre précédent</ref>...

1. Ce modèle est aussi connu sous le nom de *Chicago pizza*, par analogie avec la préparation des pizzas: ces dernières ont des ingrédients en commun, comme la sauce tomate (*core*), mais aussi des garnitures optionnelles (*additional*); elles peuvent également être apprêtées de différentes façons, comme par exemple en *calzone* (*base*).

– Le *contenu* est donné par tout élément du texte appartenant ou qui se rattache au discours. Il peut s'agir du texte originel du document, mais aussi d'informations ajoutées lors de l'encodage pour le compléter ou le paraphraser (*supplied content*). Ainsi, dans l'exemple ci-dessous, l'abréviation «*RI*» et son expansion «*Recherche d'Informations*» sont tous deux des éléments du contenu.

<abbr expan='Recherche d'Information'>RI</abbr>

Les *corrections* sont aussi considérées comme éléments de contenu. Dans l'exemple ci-dessous, «*Mojicans*», la chaîne erronée, et «*Mohicans*», sa correction, sont tous deux éléments du contenu.

<sic corr=Mohicans>Mojicans</sic>

– La *disposition* concerne la pagination, les numéros de lignes, l'apparence physique des caractères, etc. Exemples:

<pb n='42'> (numéro de page)
<hi rend=bold>Eureka!</hi> (texte en gras)

– La *documentation* concerne les caractéristiques extérieures au discours: elles se traduisent dans TEI par l'entête électronique, les annotations, interventions éditoriales, etc. Voici par exemple une entête électronique, qui spécifie le titre (<title>) et l'auteur (<author>) d'un document:

<titlestmt>
<title> Le dernier des Mohicans (version électronique) </title>
<author> Fenimore Cooper </author>
</titlestmt>

Ces informations s'organisent dans un document selon la grammaire TEI, qu'il serait trop long de décrire ici. Pour un exemple typique de document, voir la figure 3.15.

Bibliographie

[Ant95] Stéphanie ANTON. « Techniques d'interaction adaptées à la tâche de recherche d'informations ». rapport de recherche MRIM RAP95-003, Groupe MRIM – LGI-IMAG, 1995.

[Bar89] Jon BARWISE. *The Situation in Logic*. CSLI, 1989.

[BBG+95] David T. BARNARD, Lou BURNARD, Jean-Pierre GASPART, Lynne A. PRICE, et C.M. SPERBERG-MCQUEEN. Hierarchical encoding of text: Technical problems and SGML solutions. Dans Ide and Véronis [IV95], pages 211–231. Aussi dans *Computers and the Humanities* 29(1–3), 1995.

[BCR85] P. BOSC, M. COURANT, et S. ROBIN. « Recherche d'informations basée sur la comparaison d'objets ». Dans *Recherche d'Informations Assistée par Ordinateur (RIAO)*, pages 273–292, 1985.

[BD92] Robert BURGIN et Martin DILLON. « Improving disambiguation in FASIT ». *JASIS*, 43(2):101–114, 1992.

[Ber88] Catherine BERRUT. « *Une méthode d'indexation fondée sur l'analyse sémantique de documents spécialisés. Le prototype RIME et son application à un corpus médical.* ». Thèse de doctorat, Université Joseph Fourier, 1988.

[BH64] Yehoshua BAR-HILLEL. *Language and Information: selected essays on their theory and application*. Addison-Wesley, 1964.

[BH94] Peter D. BRUZA et Theodorus W.C. HUIBERS. « Investigating aboutness axioms using information fields ». Dans W. Bruce CROFT et C.J. van RIJSBERGEN, éditeurs, *Proceedings of 17th Annual International ACM-SIGIR, Dublin, Ireland*, pages 112–121. Springer-Verlag, juillet 1994.

[BK81] Duncan A. BUELL et Donald H. KRAFT. « Threshold Values and Boolean Retrieval Systems ». *Information Processing & Management*, 17(3):127–136, 1981.

[BM88] Ronald J. Brachman et Deborah L. McGuiness. « Knowledge representation, connectionism, and conceptual retrieval ». Dans *Proceedings of the 11th ACM-SIGIR, Grenoble*, pages 161–174, 1988.

[BMB95] Catherine Berrut, Philippe Mulhem, et Pascal Bouchon. « Modelling and Indexing Medical Images : the RIME approach ». Dans *HIM 95 (Conference Hypertext, Information Retrieval, Multimedia), Constance, Allemagne*, pages 105–115, avril 1995.

[Boo80] Abraham Bookstein. « Fuzzy Requests: An Approach to Weighted Boolean Searches ». *Journal of the American Society for Information Science*, 31(4):240–247, 1980.

[BP86] Catherine Berrut et Patrick Palmer. « Solving Grammatical Ambiguities within a Surface Syntactical Parser for Automatic Indexing ». Dans *ACM-SIGIR, Pisa*, septembre 1986.

[BSM94] Lou Burnard et C.M. Sperberg-McQueen. « *Encoding for Interchange: An Introduction to the TEI* », 1994. Tutorial on Text Encoding and Information Interchange, SIGIR '94.

[BZ92] Jacques Bouaud et Pierre Zweigenbaum. « A reconstruction of Conceptual Graphs on top of a production system ». Dans *Proceedings of the 7th annual workshop on Conceptual Graphs, Las Cruces*, juillet 8–10 1992.

[Cal94] James P. Callan. « Passage-level evidence in document retrieval ». Dans *Proceedings of the 17th ACM-SIGIR, Dublin, Ireland*, pages 302–310, 1994.

[Car83] Jean Caron. *Les régulations du discours, Psycholinguistique et pragmatique du langage*. Presses Universitaires de France, 1983.

[CB96] Jean-Pierre Chevallet et Marie-France Bruandet. A Study of System and User Relevance in Information Retrieval. Dans *Work Part 1 Deliverable 2: A Logic for Information Retrieval*, pages 81–114. ESPRIT BRA Project No. 8134 - FERMI, 1996.

[CCB94] Charles L.A. Clarke, G.V. Cormack, et F.J. Burkowski. « An algebra for Structured Text Search and a Framework for its Implementation ». Rapport Technique CS-94-30, Department of Computer Science, University of Waterloo, Canada, 1994.

[CD87] Yves Chiaramella et Bruno Defude. « A prototype of an Intelligent System for Information Retrieval: IOTA ». Dans *ACM-SIGIR, New Orleans*, juin 1987.

[CDB86] Yves Chiaramella, Bruno Defude, et Marie-France Bruandet. « IOTA: A full text information retrieval system ». Dans *Proceedings of the 9th ACM-SIGIR, Pisa, Italy*, pages 207–213, 1986.

[Che92] Jean-Pierre CHEVALLET. « *Un Modèle Logique de Recherche d'Informations appliqué au formalisme des Graphes Conceptuels. Le prototype ELEN et son expérimentation sur un corpus de composants logiciels* ». Thèse de doctorat, Université Joseph Fourier, Grenoble, 1992.

[CL87] W. Bruce CROFT et David D. LEWIS. « An approach to natural language processing for document retrieval». Dans *Proceedings of the 10th ACM SIGIR, New Orleans, LA*, pages 26–32, 1987.

[Con87] J. CONKLIN. « Hypertext: An introduction and survey ». *IEEE Computer*, 20(9):17–41, 1987.

[Dan74] F. DANEŠ. Functional sentence perspective and the organization of text. Dans F. DANEŠ, éditeur, *Papers on FSP*. Prague: Academia, 1974.

[DBS90] P. DEVANBU, R.J. BRACHMAN, et P.G. SELFRIDGE. «LaSSIE: a classification-based software information system ». Dans *Proceedings of the International Conference on Software Engineering, IEEE*, pages 249–261, 1990.

[Den94] Nathalie DENOS. « Pertinence en recherche d'informations : synthèse de l'état de l'art et perspectives ». rapport de recherche MRIM SUR94-002, Groupe MRIM, LGI-IMAG, 1994.

[Den96] Nathalie DENOS. « A conceptual model for user relevance in IR, and a first formalization». Deliverable 3, ESPRIT-BRA Project No. 8134 - FERMI, 1996.

[DG83] Martin DILLON et Ann GRAY. « Fasit: a fully automatic syntactically based indexing system ». *JASIS*, 34(2):99–108, 1983.

[Dun91] Mark DUNLOP. « *Multimedia Information Retrieval* ». Thèse de doctorat, University of Glasgow, 1991.

[Fag88] J.L. FAGAN. « *Experiments in Automatic Phrase Indexing for Document Retrieval: a Comparison of Syntactic and Non-Syntactic Methods* ». Thèse de doctorat, Cornell University, 1988.

[G+91] G.H. GONNET et OTHERS. « Lexicological indices for text: inverted files vs. PAT trees ». Rapport Technique OED-91-01, University of Waterloo, 1991.

[Gau96] Gilles GAUTHIER. « Prise en compte de l'utilisateur dans les systèmes de recherche d'information ». rapport de dea, Groupe MRIM et équipe IHM – CLIPS-IMAG, 1996.

[Gen91] Damien GENTHIAL. « *Contribution à la construction d'un système robuste d'analyse du français* ». Thèse de doctorat, Université Joseph Fourier, 1991.

[GF92] Michael R. GENESERETH et Richard E. FIKES. « Knowledge Interchange Format, version 3.0: Reference Manual». Rapport Technique Logic-92-1, Stanford University, Department of Computer Science, juin 1992.

[Gol91] Charles GOLDFARB. *The SGML Handbook.* Oxford University Press, 1991.

[Gre80] Maurice GREVISSE. *Le bon usage: Grammaire française avec des Remarques sur la langue française d'aujourd'hui.* Duculot, 1980.

[Gri71] H.P. GRICE. Meaning. Dans D.D. STEINBERG et L.A. JAKOBOVITS, éditeurs, *Semantics. An interdisciplinary reader in Philosophy, Linguistics and Psychology*, pages 53–59. Cambridge University Press, 1971.

[Gri75] H.P. GRICE. Logic in conversation. Dans P. COLE et J.L. MORGAN, éditeurs, *Syntax and Semantics. Vol. 3: Speech Acts*, pages 41–58. Academic Press, New York, 1975.

[Gru92] Thomas R. GRUBER. « Ontolingua: A Mechanism to Support Portable Ontologies ». Reference manual, Knowledge Systems Laboratory, juin 1992.

[Gru93] Thomas R. GRUBER. « A Translation Approach to Portable Ontology Specifications ». technical report KSL 92-71, Knowledge Systems Laboratory, avril 1993.

[GS86] Barbara J. GROSZ et Candace L. SIDNER. « Attention, intentions, and the structure of discourse ». *Computational Linguistics*, 12(3):175–204, juillet-septembre 1986.

[HD95] Theodorus HUIBERS et Nathalie DENOS. « A qualitative ranking method for logical information retrieval models». rapport de recherche MRIM RAP95-005, Groupe MRIM – LGI-IMAG, 1995.

[Hea94] Marti A. HEARST. «Multi-paragraph segmentation of expository texts». Rapport Technique UCB/CSD 94/790, University of California, Berkeley, janvier 1994.

[HH76] M. A. K. HALLIDAY et R. HASAN. *Cohesion in English.* London: Longman Group, 1976.

[HL93] Julia HIRSCHBERG et Diane LITMAN. « Empirical studies on the the disambiguation of cue phrases ». *Computational Linguistics*, 19(3):501–530, 1993.

[HP93] Marti A. HEARST et Christian PLAUNT. «Subtopic structuring for full-length document access». Dans *Proceedings of the 16th ACM-SIGIR, Pittsburgh, PA*, pages 59–69, 1993.

[HS84] Eva HAJIČOVÁ et Petr SGALL. From Topic and Focus of a Sentence to Linking in a Text. Dans B.G. BARA et G. GUIDA, éditeurs, *Computational Models of Natural Language Processing*, pages 151–163. Elsevier Science Publishers B. V. (North Holland), 1984.

[Ing92] Peter INGWERSEN. *Information Retrieval Interaction.* Taylor Graham, 1992.

[IV95] Nancy IDE et Jean VÉRONIS, éditeurs. *Text Encoding Initiative: Background and Context.* Kluwer Academic Publishers, 1995. Aussi dans *Computers and the Humanities* 29(1–3), 1995.

[Jon95] Karen Sparck JONES. « Reflections on TREC ». *Information Processing and Management*, 31(3):291–314, 1995.

[JR90] Paul S. JACOBS et Lisa F. RAU. «SCISOR: extracting information from online news ». *Communication of the ACM*, 33(11):88–97, 1990.

[Kay84] Daniel KAYSER. « Examen de diverses méthodes utilisées en représentation des connaissances ». Dans *Actes du colloque AFCET RF-IA, Paris*, pages 115–144, 1984.

[KC95] Ammar KHEIRBEK et Yves CHIARAMELLA. « Integrating Hypermedia and Information Retrieval with Conceptual Graphs ». Dans *HIM'95, Konstanz, Germany, April*, pages 47–60, 1995.

[Ker84] Dalila KERKOUBA. « *Une méthode d'indexation automatique des documents fondée sur l'exploitation de leurs propriétés structurelles. Application à un corpus technique.* ». Thèse de doctorat, Institut National Polytechnique de Grenoble, 1984.

[Khe95] Ammar KHEIRBEK. « *Modèle d'intégration d'un système de recherche d'informations et d'un système hypermédia basé sur le formalisme des graphes conceptuels. Application au système RIME.* ». Thèse de doctorat, Université Joseph Fourier, 1995.

[LC90] David D. LEWIS et Bruce W. CROFT. «Term clustering of Syntactic Phrases». Rapport Technique COINS 90-71, University of Massachusetts, 1990.

[Len95] D.B. LENAT. «CYC: A Large-Scale Investment in Knowledge Infrastructure». *Communications of the ACM*, 38(11), 1995.

[Loe94] Arjan LOEFFEN. « Text Databases: A Survey of Text Models and Systems ». *SIGMOD Record*, 23(1):97–106, mars 1994.

[Mau91] Michael L. MAULDIN. « Retrieval performance in FERRET: A conceptual information retrieval system ». Dans *Proceedings of ACM-SIGIR, Chicago, IL*, pages 347–355, 1991.

[Mec95] Mourad MECHKOUR. « *EMIR2. Un Mode'le e'tendu de repre'sentation et de correspondance d'images pour la recherche d'informations. Application a' un corpus d'images historiques.* ». Thèse de doctorat, Université Joseph Fourier, 1995.

[Mel87] Alan K. MELBY. « Statutory Analysis ». Rapport Technique, Birgham Young University, Provo, Utah, octobre 1987.

[MF96] Philippe MULHEM et Franck FOUREL. A Way to Model Multimedia Structured Documents. Dans *Work Part 3 Deliverable 4*. ESPRIT BRA Project No. 8134 - FERMI, 1996.

[MH91] Jane MORRIS et Graeme HIRST. « Lexical cohesion computed by thesaural relations as an indicator of the structure of text ». *Computational Linguistics*, 17:21–48, 1991.

[MHCW89] D.P. METZLER, S.W. HAAS, C.L. COSIC, et L.H. WHEELER. « Constituent object parsing for information retrieval and similar text processing problems ». *JASIS*, 40(6):398–423, 1989.

[Mil96] George A. MILLER. « WORDNET: A Lexical Database for English ». *Communications of the ACM*, 38(11):39–41, 1996.

[MMT89] William C. MANN, Christian M.I.M. MATTHIESSEN, et Sandra A. THOMPSON. « Rhetorical Structure Theory and Text Analysis ». Rapport Technique ISI/RR-89-242, Information Science Institute, 1989.

[MN96] Philippe MULHEM et Laurence NIGAY. « Interactive Information Retrieval Systems: From User Centred Interface Design to Software Design ». Dans *19th ACM SIGIR Conference, Zurich, Suisse*, pages 326–334, 1996.

[Nie90] Jianyun NIE. « *Un modèle logique général pour les Systèmes de Recherche d'Informations. Application au prototype RIME.* ». Thèse de doctorat, Université Joseph Fourier, 1990.

[NK91] S.T. NEWCOMB et N.A. KIPP. « HyTime: Hypermedia/Time-based document structuring language ». *Communications of the ACM*, 34(11):67–83, novembre 1991.

[Pai93] Hans PAIJMANS. « Comparing the document representations of two IR-systems: CLARIT and TOPIC ». *JASIS*, 44(7):383–392, 1993.

[Pal90] Patrick PALMER. « *Etude d'un analyseur de surface de la langue naturelle: application à l'indexation automatique de textes* ». Thèse de doctorat, Université Joseph Fourier, 1990.

[Par93] François PARADIS. « Indexation sémantique de documents: Application a un corpus technique ». Thèse de maîtrise, Université de Montréal, 1993.

[Par94a] François PARADIS. « A Model for Structured Textual Documents ». Technical Report Series Fermi 4/94, ESPRIT BRA Project N.8134: FERMI, 1994.

[Par94b] François PARADIS. « Revue de formalismes pour la représentation de la connaissance ». rapport de recherche MRIM SUR94-001, Groupe MRIM – LGI-IMAG, 1994.

[Par95a] François PARADIS. Modeling Textual Information. Dans *Work Part 3 Deliverable: A Model for the Semantic Content of Multimedia Data*, pages 11–67. ESPRIT BRA Project No. 8134 - FERMI, 1995.

[Par95b] François PARADIS. « Using Linguistic and Discourse Structures for Topic Indexation ». Dans *Third Natural Language Processing Pacific Rim Symposium, Seoul, Korea*, pages 157–162, décembre 1995.

[Par95c] François PARADIS. « Using Linguistic and Discourse Structures to Derive Topics ». Dans *Fourth International Conference on Information and Knowledge Management (CIKM), Baltimore, Maryland.*, pages 44–49, novembre 29 – décembre 2 1995.

[PB96] François PARADIS et Catherine BERRUT. « Experiments with Theme Extraction in Explanatory Texts ». Dans *Second International Conference on Conceptions of Library and Information (CoLIS 2), Copenhagen, Denmark*, octobre 13–16 1996.

[Por80] M.F. PORTER. « An algorithm for suffix stripping ». *Program*, 14(3):130–137, juillet 1980.

[Rén84] Alfréd RÉNYI. *A Diary on Information Theory*. John Wiley & Sons, 1984.

[RHB92] S.E. ROBERTSON et M.M. HANCOCK-BEAULIEU. « On the evaluation of IR systems ». *Information Processing & Management*, 28(4):457–466, 1992.

[SAB93] Gerard SALTON, James ALLAN, et Chris BUCKLEY. « Approaches to passage retrieval in full text information systems ». Dans *Proceedings of the 16th ACM-SIGIR, Pittsburgh, PA*, pages 49–58, 1993.

[Sav94] Jacques SAVOY. « A Learning Scheme for Information Retrieval in Hypertext ». *Information Processing and Management*, 30(4):515–533, 1994.

[SDV95] Patrick SAINT-DIZIER et Evelyne VIEGAS. *Computational Lexical Semantics*. Cambridge University Press, 1995.

[SM83] Gerard SALTON et M.J. MCGILL. *Introduction to modern Information Retrieval*. Mcgraw Hill Book Company, New York, 1983.

[SM95a] C.M. SPERBERG-MCQUEEN. « Bare bones TEI ». *Text Technology (special issue)*, pages 248–265, 1995.

[SM95b] C.M. SPERBERG-MCQUEEN. The design of the TEI encoding scheme. Dans Ide and Véronis [IV95], pages 17–39. Aussi dans *Computers and the Humanities* 29(1–3), 1995.

[Sma94] M. SMAÏL. « *Raisonnement à base de cas pour une recherche évolutive d'informations ; prototype Cabri-n – Vers la définition d'un cadre d'acquisition des connaissances* ». Thèse de doctorat, Universit é Henri Poincar é – Nancy, France, 1994.

[Sme92] Alan F. SMEATON. « Progress in the Application of Natural Language Processing to Information Retrieval Tasks ». *The Computer Journal*, 35(3):268–278, 1992.

[ST92] Airi SALMINEN et Frank Wm. TOMPA. « PAT expressions – An algebra for text search ». Rapport Technique OED-92-02, UW Centre for New Oxford English Dictionary, University of Waterloo, 1992.

[Swa94] Michael SWAN. *Practical English Usage*. Oxford University Press, 1994.

[TAAC87] Richard M. TONG, Lee A. APPELBAUM, Victor N. ASKMAN, et James F. CUNNINGHAM. « Conceptual Information Retrieval using RUBRIC ». Dans *Proceedings of the 10th ACM-SIGIR, New Orleans*, pages 247–253, 1987.

[TEI94] TEI. « *TEI Guidelines for Electronic Text Encoding and Interchange (P3)* », 1994.

[TSM91] Jean TAGUE, Airi SALMINEN, et Charles MCCLELLAN. « Complete formal model for Information Retrieval Systems ». Dans *Proceedings of the 14th ACM-SIGIR*, pages 14–20, 1991.

[vR79] C.J. van RIJSBERGEN. *Information Retrieval*. Second Edition, Buttersworth Longon England, 1979.

[vR86] C.J. van RIJSBERGEN. « A non-classical logic for Information Retrieval ». *Computer Journal*, 29(6), 1986.

[vRW86] H. van RIEMSDIJK et E. WILLIAMS. *Introduction to the Theory of Grammars*. Cambridge: The MIT Press, 1986.

[WB88] Tony WILLIAMS et Brian BAINBRIDGE. Rule based systems. Dans *Approaches to Knowledge Representation: An Introduction*, pages 101–115. Research Studies Press Ltd., 1988.

[Wil94] Ross WILKINSON. « Effective retrieval of structured documents ». Dans *Proceedings of the 17th ACM-SIGIR*, pages 311–317, Dublin, Ireland, 1994.

[WK79] W.G. WALLER et D.H. KRAFT. « A Mathematical Model of a Weighted Boolean Retrieval System ». *Information Processing & Management*, 15(5):235–245, 1979.

[Woo75] William A. WOODS. What's in a link: foundations for semantic networks. Dans Daniel G. BOBROW et Allan COLLINS, éditeurs, *Representation and Understanding: Studies in Cognitive Science*, pages 35–82. Academic Press, New York, 1975.

[WS75] Warren WEAVER et Claude E. SHANNON. *Théorie mathématique de la communication*. Les Classiques des sciences humaines, 1975. Paru originellement sous le titre "The mathematical theory of communication" (1949).

[Yao95] Y.Y. YAO. « Measuring retrieval effectiveness based on user preference on documents ». *Journal of the American Society for Information Science*, 46(2):133–145, 1995.

Index des citations

Index thématique

– D –

– E –

– F –

– G –

– H –

– I –

– J –

– K –

– L –

– M –

– N –

– O –

– P –

– Q –

– R –

– S –

– T –

– U –

– V –

– W –